PB-229 057

STATISTICAL METHODS IN AIR POLLUTION MEDICAL STUDIES

John R. Goldsmith

California State Department of Public Health
Berkeley, California

30 April 1961

DISTRIBUTED BY:

NTIS
National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

PB 229 057

Statistical Methods in Air Pollution Medical Studies

FINAL REPORT

Contract Number SAph 73596

Prepared for:

U.S. Public Health Service
Department of Health, Education and Welfare
Division of Air Pollution

State of California
Department of Public Health
Bureau of Chronic Diseases
Berkeley, California

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

126

NOTES AND ERRATA

Page 14, the title of this section suggests that admissions to the Los Angeles County Hospital were surveyed; this is incorrect. The survey was of admissions to hospitals in Los Angeles County.

Page 14, paragraph 3, refers to Appendix I which follows page 92.

Page 37, the categories included in Table XIV are defined more extensively in Appendix II.

Page 64, last line, paragraph 3, the word "air" should be "art."

Page 82, under section 3, there is a duplication of words on all of line 5, and half of line 6.

Page 91, last line should read "there are no new variables remaining to be added."

---FINAL REPORT---

ON: Contract Number SAph 73596

TO: Department of Health, Education and Welfare
U.S. Public Health Service
Division of Air Pollution
Bureau of State Services

PURPOSE OF STUDY: To investigate the applicability of suitable statistical methodology and to develop the accompanying data reduction programs for examining the relationships between atmospheric pollution, meteorologic conditions and mortality and morbidity in Los Angeles.

CONTRACTOR: State of California
Department of Public Health
2151 Berkeley Way
Berkeley 4, California

PROJECT DIRECTOR: John R. Goldsmith, M.D.
Bureau of Chronic Diseases
California State Department of Public Health

AMOUNT: $39,160.00

PERIOD: November 1, 1959 to June 30, 1960
(Amended to extend date to April 30, 1961)

Table of Contents*

Page

Section I 3

Introduction 4

Section II 5

Summary of Previous Progress Reports 6

Section III 8

Report of Operations and Analyses

- A. Nursing Home Study 10
- B. Los Angeles County Hospital Admissions Survey 14
- C. Kaiser Foundation Health Plan Study 24
- D. Framingham Data 34
- E. Sawtelle Veterans Domiciliary Data 38

Section IV 46

Categories of Problems Which Have Been Recognized and Approaches to their Solution

- A. Multivariate Problems 48
- B. Time-Series Problems 51
- C. Air Pollution Area Coding Problems 53
- D. Systems Problems 59

Section V 61

Statistical Methods in Air Pollution Medical Studies; A Review and Appraisal . . . 62

---

*Tables and Figures referred to in the text are found at the end of each section.

Section I
Introduction

This comprehensive final report is submitted in accordance with Article 10-B of Contract SAph 73596. This report does not reiterate the contents or appendices of Progress Reports I or II. However, the contents of these reports will be summarized as required in Article I.A.-4.

At least three publications are contemplated, based, in part, upon the work done under this contract. One article, "Air Pollution and Daily Mortality," has been published in the American Journal of the Medical Sciences, 241:581 (May 1961). Another article, "Tests of Maximal Expiratory Flow in a Population Sample: An Application of Component Analysis," is being prepared for submission to the American Review of Respiratory Diseases. A third publication, "Statistical Methods for Air Pollution Medical Studies," is planned. Possibly, another may be developed on the relationships of symptoms and signs of respiratory diseases to smoking and to tests of maximal respiratory flow.

The application for a research grant on "Time-Series Methods" has been favorably acted upon and plans are underway to staff this research program.

This report is divided into five sections, including: first, this introduction. The second summarizes the first two progress reports. The third deals with the data processing operations and analytic work undertaken during the latter portion of the contract period. The fourth deals with four general problems which have been recognized during the period of this contract, three of which we have attacked: time-series problems, multivariate problems, and problems of assigning events to geographic areas. The fourth general problem is the systems analysis problem. The final section contains a critique of statistical methods which are or may be useful for the analysis of data concerning health effects of air pollution and a discussion of procedures which may be useful in the future.

Section II
Summary of Previous Progress Reports

Summary of Progress Report I

The first progress report described available data concerning mortality and/or morbidity which might be related to air pollution. Sets of data covered were:

a. General population mortality by area, by cause and by age

b. Nursing Home mortality and transfers to hospitals

c. Morbidity data from the California Health Survey

d. Morbidity data represented by hospital admissions
   1. Kaiser Foundation Health Plan July-December 1958
   2. Los Angeles County Medical Association--Southern California Hospital Council, 1958

e. Morbidity indicated by requests for home medical attention and walk-in clinic visits--Kaiser Foundation Health Plan

f. Morbidity indicated by physiological impairment
   1. Highland-Alameda County Hospital
   2. Framingham (a control population)
   3. Sawtelle Screening and Clinic

g. Morbidity indicated by alteration in clinical condition (diary)

Summary of Progress Report II

The second progress report described in detail the data to be tabulated and the inherent limitations for use of available data. It reviewed methods for, and described results of removing the variability due to sex and physique from data on lung function. In addition, a section of the report described the nature of the cyclical seasonal pattern in Los Angeles for temperature, oxidant, carbon monoxide, and cardio-respiratory mortality. A computer program for the analysis of cyclic regressions was developed later in the contract period. Finally, plans were described for further work--some of which has been accomplished and will be reported.

Sets of computer programs, the research grant draft, and an outline of data and analyses performed were appended.

Section III
Report of Operations and Analyses

Introduction

This section consists of a description of the major specific projects undertaken as part of the contract. In general, for each project the source of the data is described, along with its limitations, and the data processing and analytical methods used. Comments are made on problems encountered, and further analysis which might be undertaken is suggested.

A. A Study of Mortality and Morbidity in Nursing Homes

Since July 1, 1955 the State Health Department has received weekly reports of deaths and transfers to hospitals from a group of nursing homes in Los Angeles County. These data were examined on a current day-to-day basis as a means of surveillance of the health effects of air pollution on a relatively sensitive population. The work to be reported is a general descriptive summary of the data collected to date. Detailed statistical analysis is underway.

Source of Data. All Los Angeles County Nursing Homes with at least 25 beds as of July 1, 1955 were asked to participate. The study group was enlarged on January 1, 1958 to include an additional group of homes meeting this requirement. These were homes which either had increased their capacity or had been established since the beginning of the study. Reports were received from a total of 98 nursing homes, 13 of these being in the "new group" admitted in 1958. The total bed capacity of the homes varied from 3690 to 4729 during the four and one-half year period for which data are available.

Description of Data. The items of information requested from the nursing homes are: name, date of death or transfer to hospital, date of admission to the nursing home, age and sex. Cause of death is not included but could be obtained from the death certificates of individuals who died. Reports are requested on a weekly basis; however, some homes accumulate data over a longer time period before submitting them. Upon receipt, reports are edited for completeness, plausibility, and duplication of reported data.

It should be emphasized that very little is known about the population base, that is, the characteristics of the people in nursing homes, beyond the probable size of the population and a very general description in terms of age, sex distribution, and health status. It is assumed of necessity that the population characteristics of the occupants of the homes as a whole remain relatively constant over time. If comparisons are made between homes in different geographic areas, however, the possibility of differences in age, socio-economic status, or type and severity of illness must be considered. Comparisons of this kind are discussed later in the report as part of the general problem of coding by geographic area.

Table I shows the volume of data gathered and its distribution by sex. Since the bed capacity of the homes is constantly changing, comparisons over time require the use of rates. Table II shows the average weekly rates of deaths and transfers to hospitals by year. A constant percent of occupancy of total beds per home is assumed.

Data Processing. As mentioned previously, reports were edited by hand as they were received. Nursing homes were coded by geographic location using three different coding schemes, two serving primarily to provide comparability with other data, and one using the natural geographic groupings of the homes. These are described in Section IV-c, "Air Pollution Area Coding Problems." As an aid in evaluating the surveillance aspect of the reporting system, the date of receipt of the report was coded for 1959 data. Following completion of punching, additional editing and checking were done. Preliminary tabulations were then obtained, including rates and percentage distribution of some data.

In view of the possibility of other similar studies being done, one observation is significant with respect to the computation of rates. The number of beds included in the study changed every few weeks because of changes in the licensed bed capacity of individual homes. Denominators for the original and for the new group of nursing homes were calculated by week. This information and the weekly counts of deaths and transfers to hospitals were then used for computation of weekly rates by electronic calculator. This method was used because a more detailed time breakdown, that is, by day or by geographic location of home would have resulted in very small numerators. In general, however, it would probably be more efficient to transfer to tape the most detailed denominator data available, that is, number of beds for each home for each day. Calculation of rates could then be made by computer for any possible combination of homes for any time breakdown without additional hand calculation of denominators.

Tabulations Available. General descriptive tabulations of deaths and transfers for each year were obtained by nursing home, by age, sex, and year, and by length of stay in nursing home. Time-series tabulations were run by day for old and new homes as groups, and by week for homes in specified geographic areas. Table III summarizes the age distribution of the deaths and transfers. The basic age tabulations contain greater detail.

Further Analysis Planned. Regression analysis of time-series data is planned, possibly fitting a Fourier series and using residuals as the dependent variable if the data appear to be suitable. Periods of high and low air pollution will be compared. Cause of death may be examined for selected time periods. Data will also be examined from the viewpoint of evaluating the accuracy and efficiency of the system of data collection and processing.

Table I

LOS ANGELES COUNTY NURSING HOME STUDY
DEATHS AND TRANSFERS TO HOSPITALS BY SEX
JULY, 1955-DECEMBER, 1959

YEAR	DEATHS			TRANSFERS TO HOSPITALS		
	Total	Male	Female	Total	Male	Female
Total	10,622	3,825	6,793	4,669	1,596	3,070
1955 (July-December)	1,126	373	753	523	174	349
1956	2,103	722	1,381	1,043	374	668
1957	2,109	772	1,334	977	319	656
1958	2,698	993	1,704	1,099	343	756
1959	2,586	965	1,621	1,027	386	641

Note: Totals include 7 sex not stated.
Bed capacity of the nursing homes varied from 3,690 to 4,729 during this period.

Table II

LOS ANGELES COUNTY NURSING HOME STUDY
AVERAGE WEEKLY DEATH AND HOSPITAL TRANSFER RATES
JULY, 1955-DECEMBER, 1959

YEAR	DEATHS	TRANSFERS TO HOSPITALS
1955 (July-December)	11.5	5.4
1956	10.8	5.3
1957	10.4	4.8
1958	11.3	4.6
1959	10.6	4.2

Note: Rates are per 1,000 nursing home beds.

Table III

LOS ANGELES COUNTY NURSING HOME STUDY
PERCENTAGE AGE DISTRIBUTION OF DEATHS AND TRANSFERS TO HOSPITALS BY SEX
JULY, 1955-DECEMBER, 1959

AGE	DEATHS			TRANSFERS		
	Total	Male	Female	Total	Male	Female
All Ages	100.0	100.0	100.0	100.0	100.0	100.0
Less than 65	8.1	9.2	7.4	8.9	11.3	7.7
65-74	20.9	24.0	19.2	20.4	23.3	18.9
75-84	41.5	41.1	41.8	46.7	44.4	47.9
85-94	26.6	23.3	28.5	22.1	18.9	23.7
95 and over	2.9	2.4	3.1	1.9	2.1	1.8

Source: Nursing Home Records.

B. Los Angeles County Hospital Admissions Survey

Morbidity data is, in general, sparse and usually incomplete. When the Los Angeles County Medical Association, in conjunction with the Hospital Council of Southern California, suggested studying hospital admissions for illnesses which may be caused or aggravated by exposure to air pollutants, the State Department of Public Health felt this was an opportunity to advance our knowledge in this area.

These data represent morbidity of a relatively serious nature since hospitalization is the criterion for inclusion in the study. Furthermore, hospitalization in a county-supported institution, due in part to bed limitations, would probably be for relatively more serious illnesses than might be observed in private hospitals. This possibility will be investigated in later analysis. In the private hospitals some socioeconomic selection probably occurs because the individuals must either be economically able to pay for hospitalization or have insurance which will cover such hospitalization.

Source of Data. The data come from the medical records libraries of 35 hospitals in the Los Angeles Metropolitan Area. Medical Records Librarians in each of the participating hospitals were requested to abstract selected data from the hospital charts of patients admitted during 1958 with specific cardiac and/or respiratory discharge diagnoses (See Appendix I). The discharge diagnostic groups selected were those which possibly might be related to, or which might be aggravated by air pollutants.

The librarians were paid for each chart abstracted to insure the inclusion of all hospital charts with the selected discharge diagnoses.

In some cases, unsolicited discharge diagnoses were included in the abstracts. This is mentioned because it points out the need for careful editing prior to payment when data are obtained in this manner. Personal supervision and guidance of the librarians would reduce the possibility of this type of error and would also result in increased uniformity in the abstracted entries. The lack of uniformity obtained made line-by-line, entry-by-entry editing a necessity which needlessly increased the work. This increased cost would have netted better results if it had been invested in training rather than in repairing. Ideally, such abstracts should, with a minimum of spot checking, be immediately ready for the key-punch operator.

Description of Data. Besides the discharge diagnosis (or diagnoses), twelve other items of information were collected for each patient. They are as follows:

1. Disposition: discharged, expired
2. Hospital chart number
3. Date of admission: month, day, year (all admisision occurring between January 1, 1958 and December 31, 1958 for selected causes)
4. Sex
5. Age
6. Health insurance coverage
7. Number of days in hospital

*8. Home area (1-6)

*9. Employment area (1-6, other, don't know)

10. Onset of this illness: month, day, year
11. Hemoglobin at time of admission
12. Hematocrit at time of admission

---

*The home and employment areas refer to six geographical areas which have a monitoring station for several air pollutants. These data permit the estimation of the probable exposure experienced by an individual in our sample. Inclusion in this study is dependent upon the individual's living in one of the six areas.

Data Processing. One of the first problems was that of multiple diagnoses. Since it is not uncommon for a single patient to have more than one of the selected discharge diagnoses, allowance was made for the recording of two cardiac and two respiratory diagnoses. In the event that more than two selected cardiac and/or respiratory conditions were recorded for a single patient, this fact was coded and punched on the card.

After punching, the cards were again edited for errors and inconsistencies. Some machine coding was done to facilitate tabulating. The Standard Nomenclature Diagnoses Codes used by the hospitals are seven and eight digit numbers. Such large numbers are cumbersome; also, certain of the diagnoses were very closely related. Thus, on the basis of these relationships, eight cardiac and fourteen respiratory disease groups were designated. This is shown in Appendix I. Special two-digit codes were assigned to these groups and gang-punched on the cards. Descriptive tabulations were then obtained.

There is more machine editing and coding to be done, such as the calculation of the number of days between onset and admission to the hospital and the assignment of day-of-week and week-of year codes to day of onset and admission.

Tabulations Available. A number of descriptive tabulations were obtained to give some idea of the nature of the distributions. At present, tabulations are available stratified by discharge diagnosis by: 1) day of admission, 2) day of onset, 3) age and sex, 4) home and employment areas. Additional analysis is being planned from these tabulations.

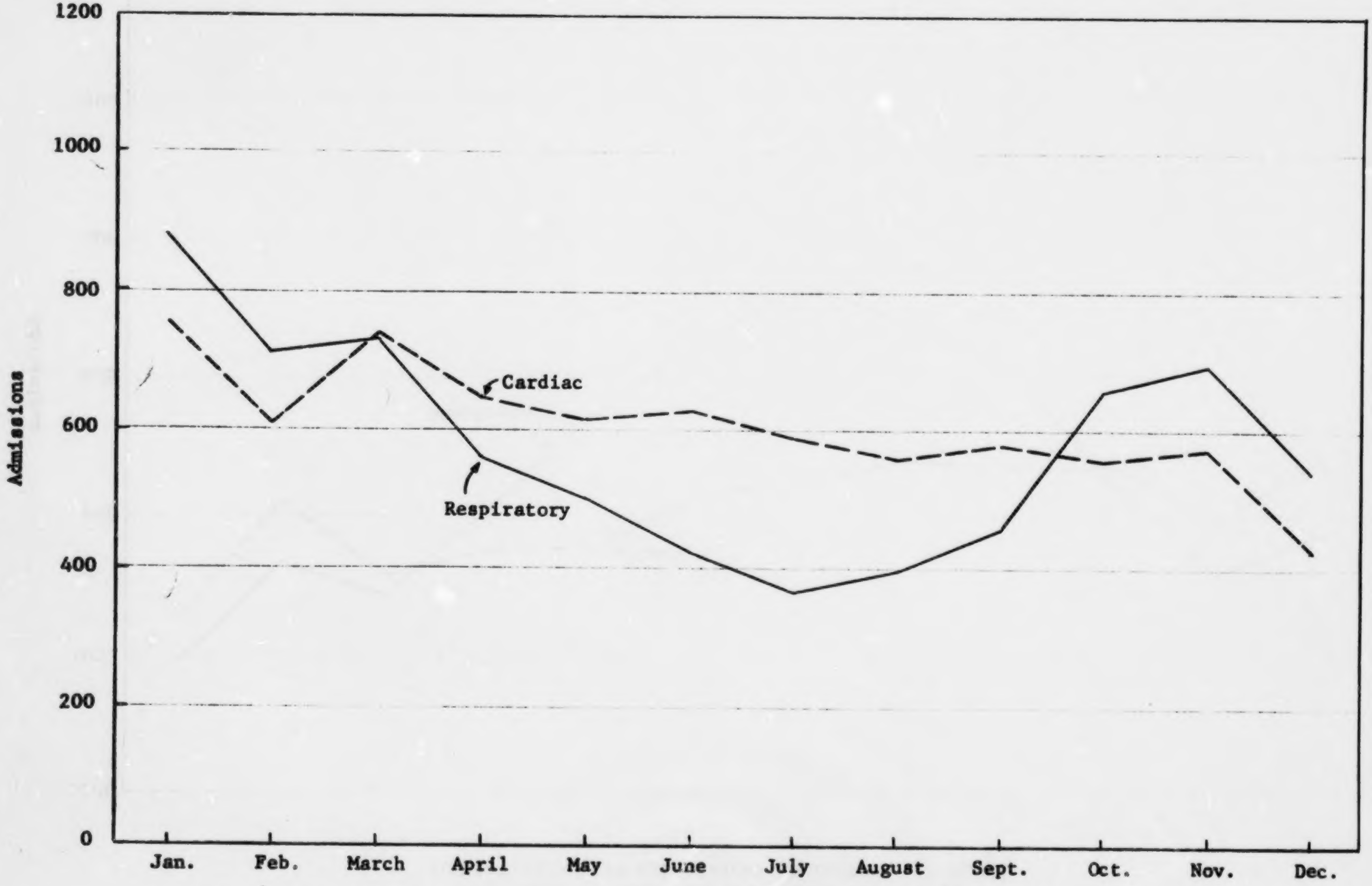
Preliminary Analysis. The admissions by day for both cardiac and respiratory diagnoses are so erratic that any trend or cycle that may be present is obscured by the daily variations, but the monthly totals do show a definite cyclic pattern, as expected. The respiratory admissions and onsets dropped to a low in July and the cardiac admissions and onsets dropped in August. However, these numbers are not adjusted for length of calendar month; therefore the drop in both cardiac and respiratory admissions during the month of February is due in part, at least, to the shorter period of observation, that is, 28 days, rather than 30 or 31.

When onsets by day are graphed, (Figures I and II), there are marked peaks on the first day of each month. There are also regular rises for days with numbers divisible by five. This is a common problem of recall by the patient and poses definite analytic problems. Grouping onsets by week would not have solved this problem because some weeks will contain two such peaks, while others will contain only one. Limitations of this nature can not be overlooked.

Scatter diagrams, Figures III to VI, give a picture of the day-to-day relationship of respiratory and onsets admissions to basin averages for carbon monoxide and oxidant. The basin average is defined as the mean of the maximum hourly averages for five stations located in widely separated areas of Los Angeles County. No attempt was made to weight these various readings by the populations at risk surrounding each of the five stations.

Figure I

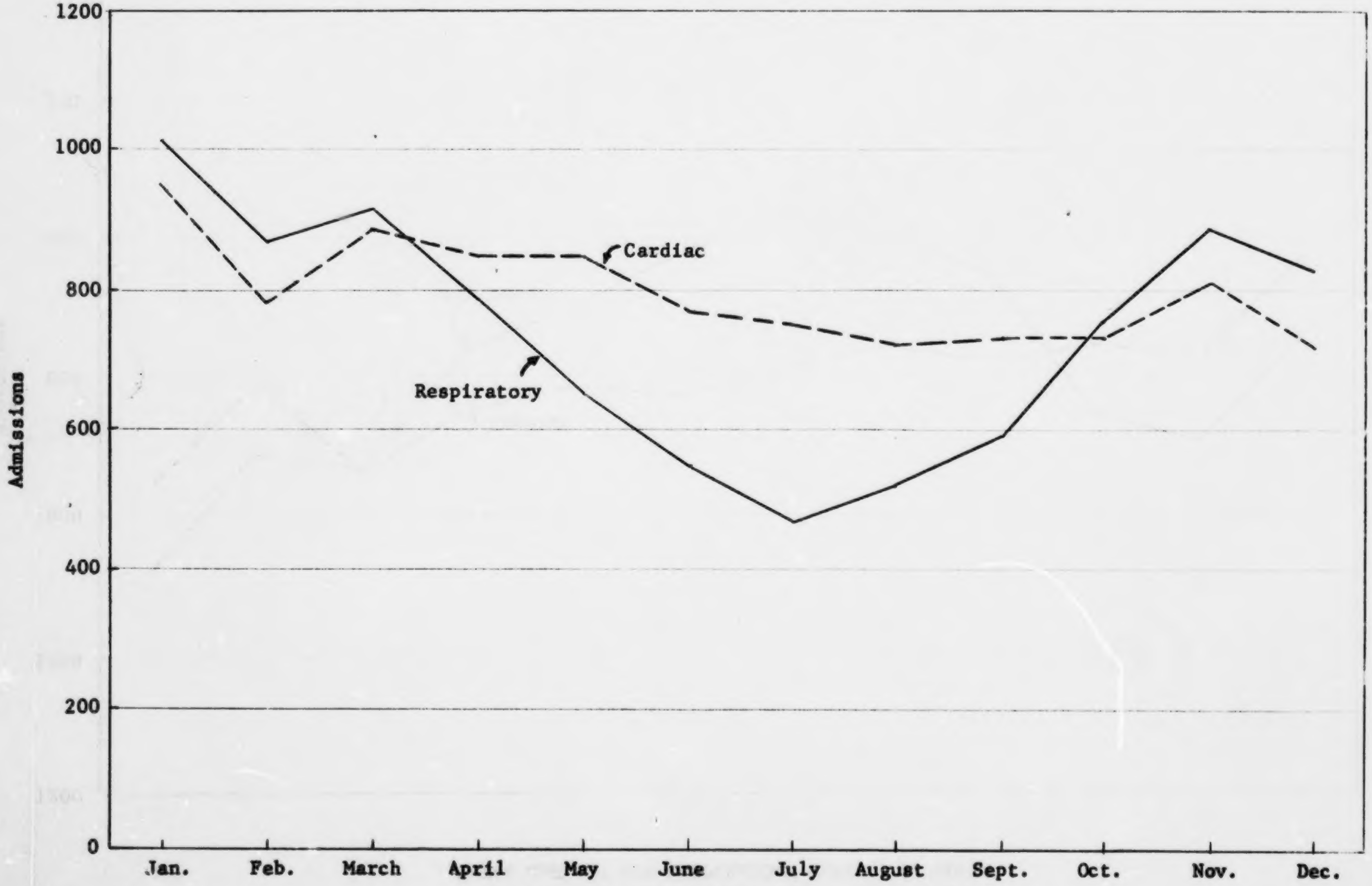
LOS ANGELES HOSPITAL ADMISSIONS STUDY 1958
TOTAL CARDIAC AND RESPIRATORY ONSETS, BY MONTH



Source: State of California, Department of Public Health, Air Pollution Medical Studies.

Figure II

LOS ANGELES HOSPITAL ADMISSIONS STUDY 1958
TOTAL CARDIAC AND RESPIRATORY ADMISSIONS, BY MONTH



Source: State of California, Department of Public Health, Air Pollution Medical Studies.

Figure III

LOS ANGELES HOSPITAL ADMISSIONS STUDY
TOTAL RESPIRATORY ONSETS BY OXIDANT
January through December, 1958

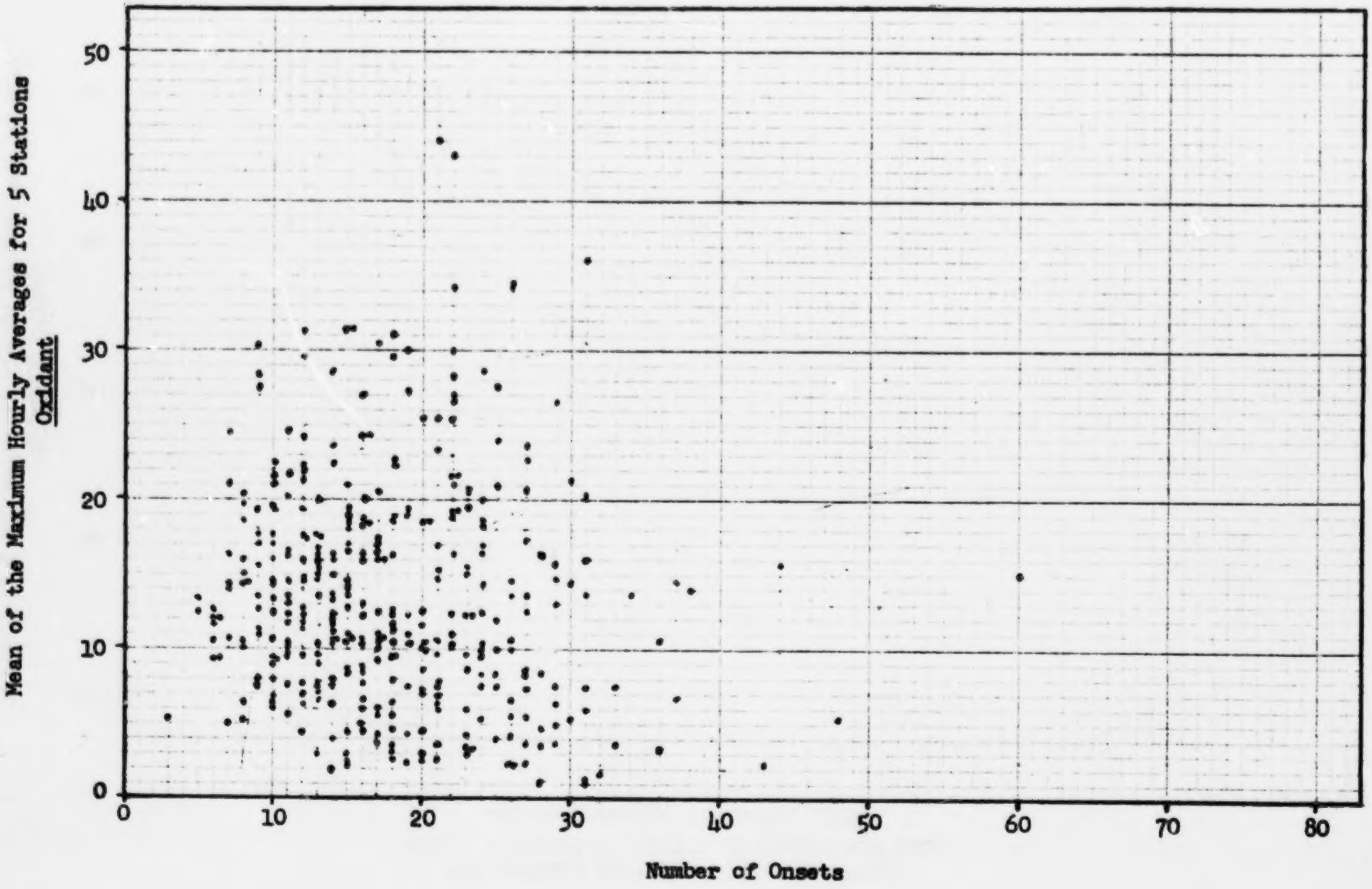


Figure IV

LOS ANGELES HOSPITAL ADMISSIONS STUDY
TOTAL RESPIRATORY ONSETS BY CARBON MONOXIDE

January through December, 1958

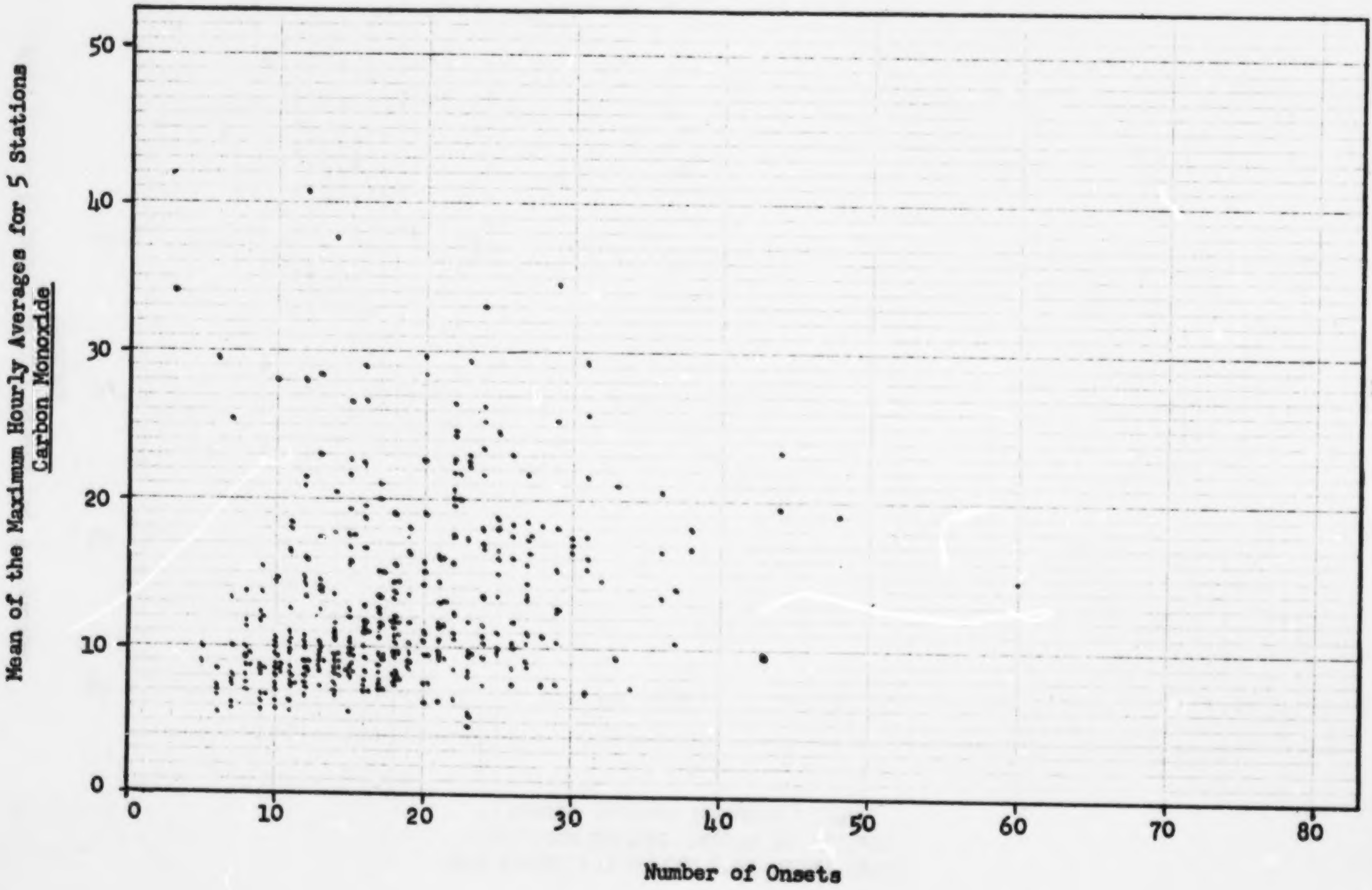


Figure V

LOS ANGELES HOSPITAL ADMISSIONS STUDY
TOTAL RESPIRATORY ADMISSIONS BY OXIDANT
January through December, 1958

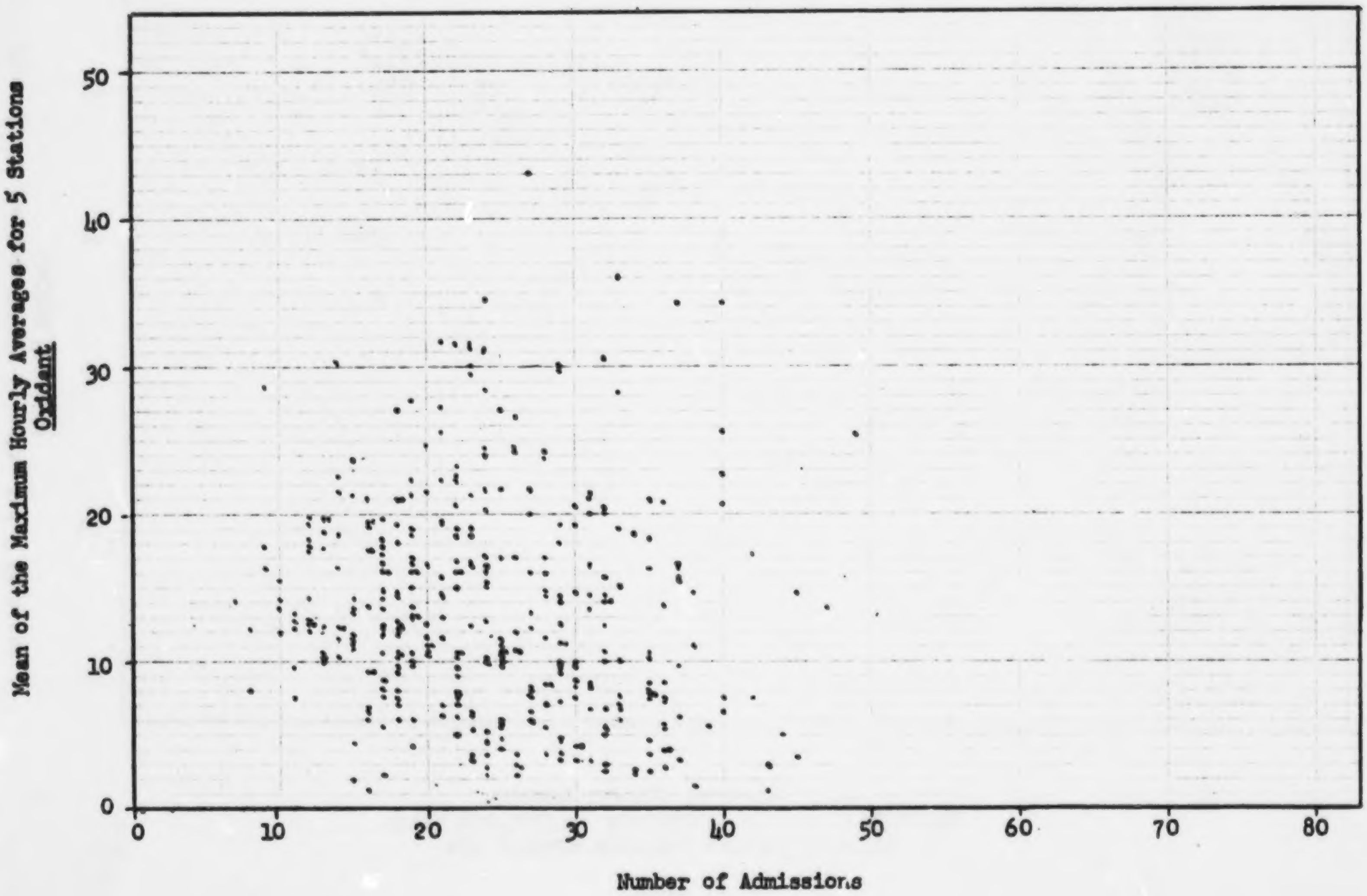
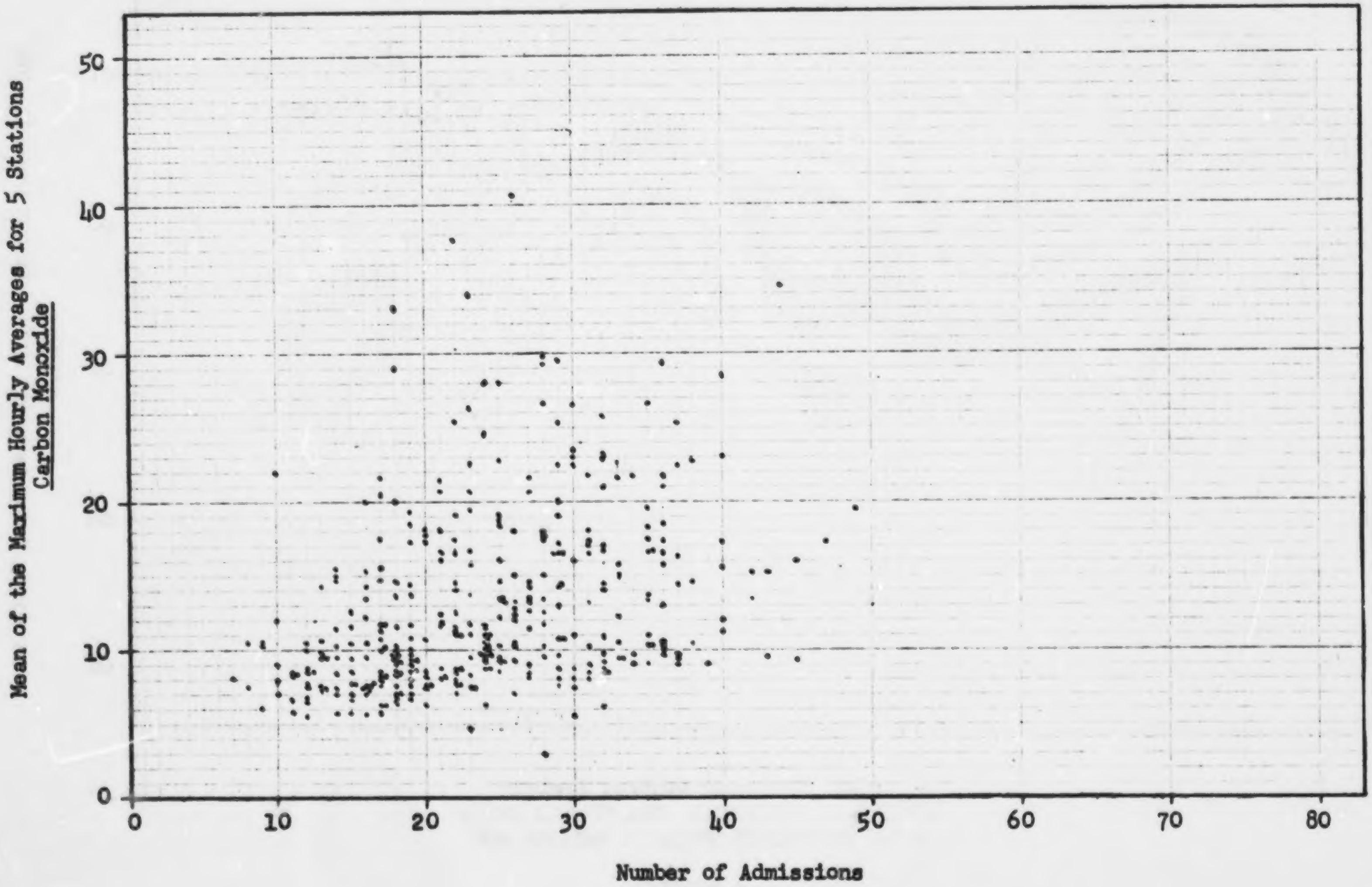


Figure VI

LOS ANGELES HOSPITAL ADMISSIONS STUDY
TOTAL RESPIRATORY ADMISSIONS BY CARBON MONOXIDE
January through December, 1958



C. Kaiser Foundation Health Plan Study

Mortality studies of the acute health effects of air pollution have been critized for the following reasons:

1. Death is only one of many possible health effects of air pollution.
2. Death, in the absence of a disaster, is probably a rare acute effect of air pollution.
3. Death due to air pollution is more apt to be a chronic effect than an acute effect.
4. Death as an effect, whether acute or chronic, is more apt to occur among the aged or the chronically ill than among the general population.

Studies of morbidity have been hampered by lack of suitable data. However, it was suggested that the records of the Kaiser Foundation Health Plan might yield data which would serve as an indicator of the acute effects of air pollution on morbidity. The present study was undertaken to determine whether data of this type are appropriate for analysis. It should be pointed out that only acute effects, that is, effects occurring on the same day or within several days of exposure to air pollution are being studied.

Three sets of data covering the period from July 1 through December 31, 1958 were obtained from the Southern California Region of the Kaiser Foundation Health Plan. They included admissions to Fontana, Harbor City, and Sunset Hospitals; walk-in visits to clinics at five locations in the region; and home calls throughout the region.

As an initial step in analysis, each set of data was tabulated by day. Tables IV through XI summarize these data by month. The detailed tabulations were useful in showing the magnitude of the daily numbers and the amount of day-to-day variation. Punch cards summarizing the data by day were obtained at the same time as the tabulations as an aid in further analysis.

Daily admissions to Fontana and Harbor hospitals proved to be too small, especially when surgical and gynecological admissions were omitted, to be useful as a means of relating day-to-day changes in morbidity to daily variation in weather and air pollution. Consequently, analysis of hospital admissions was limited to data from Sunset Hospital. Walk-in clinic visits and house calls proved to be of sufficient magnitude for analysis, but presented other problems.

Daily tabulations were examined for obvious cyclic and trend effects and for obvious relationships with temperature, total oxidant and carbon monoxide. All sets of data showed an apparent day-of-week effect. Since some environmental variables also show a day-of-week effect, this factor will require further analysis. Also, all sets of data showed a general increase with time. This may be a combined effect of an increase in health plan membership and a general increase in morbidity as the winter months approach. No obvious relationships with environmental measurements were seen.

In order to detect departures from "chance" variation of the data as a whole, Poisson distributions were fitted to frequency distributions of the number of Sunset Hospital admissions per day in pediatrics and internal medicine. (Figures VII and VIII). The results were negative or inconclusive.

House calls were chosen for regression analysis because they offered appropriate, although not necessarily accurate, diagnostic categories. The ratio of cardio-respiratory complaints to total house calls was used as the dependent variable in order to eliminate some of the trend effect. The independent variables used to represent air pollution were total oxidant and carbon monoxide. Temperature was included as an independent variable because of its known effect upon mortality and its relationship to air pollution. The computations were done on the IBM 704, using a program provided by the Department of Biostatistics, University of California at Los Angeles. Table X shows the simple correlation coefficients obtained. In general, the correlations were small with the exception of that between total oxidant and temperature. The negative correlation between the dependent variable and temperature and the positive correlation between the dependent variable and carbon monoxide may result from seasonal relationships. Table XI summarizes the results of the analysis. The F-Values are all significant at the one percent level. In view of the many time-related factors involved, the statistically significant results cannot be interpreted as implying causation; unfortunately, the six-month period being studied does not give adequate data for the removal of seasonal influences.

Regression analysis might profitably be repeated using a different measure of house calls as the dependent variable. Hospital admissions were felt to be less reliable than clinic visits or house calls as an indicator of the immediate day-to-day effects of air pollution, and no diagnostic information was available for walk-in clinic visits; therefore, these data were not analyzed by the regression method.

Problems of Analysis. Perhaps the greatest difficulty in analyzing data of this type is presented by the numerous variables which determine how many people will seek medical aid on a given day. First, there are the determinants of illness itself, such as infectious agents, susceptible hosts, pre-existing physical or physiological factors, and environmental factors besides those which are under study. Second, granted that illness exists, what determines whether the patient seeks medical attention or the method by which he seeks it? For example, would he choose a visit to a walk-in clinic or a house-call by a physician? Related to this are the factors which determine whether a patient will be a hospital admission.

These factors vary in two ways, first, among the population at risk (in this case the health plan membership), and second, over time. One may, perhaps, be able to assume that the population remains relatively constant, as a whole, over time in its response to causative factors of illness and its response to the presence of illness itself. On the other hand, one cannot justifiably assume that these factors themselves remain constant over time. Time-related factors may be divided into those that tend to exaggerate or simulate the relationship being studied and those that tend to obscure the relationship. An example of the former is the tendency of the atmospheric concentration of carbon monoxide and the incidence of illness to rise during the winter months. Seasonal fluctuations such as these may have a causal relationship, or they may each be a result of other seasonal influences. One method of coping with this situation is by attempting to remove, in a statistical sense, these seasonal fluctuations and by concentrating upon the relationship of the remaining day-to-day variation. As mentioned above, data in this study did not cover a sufficient period of time to identify seasonal variation.

Non-seasonal factors may also affect the relationship being studied. The second kind of time-related factors, those that tend to obscure the relationships under the study, are more apt to lead us to "inconclusive" than to "wrong" results. They may be seasonal factors which operate in opposition to seasonal causative factors, reducing their effect, or they may be factors which, though not seasonal, exhibit some other time-related cyclical pattern.

The problems of analysis which are inherent in time-series should also be mentioned. For example, the method of regression analysis used is based upon the assumption that successive observations are independent. This may not be true with respect to observations made during successive intervals of time.

These problems are further discussed in Section IV-B, "Time-Series Problems."

Table IV

KAISER FOUNDATION HEALTH PLAN STUDY
FONTANA HOSPITAL ADMISSIONS BY MONTH AND TYPE
JULY-DECEMBER, 1958

TYPE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER
Total	540	481	530	539	502	512
Medical	103	86	101	107	97	104
Surgical	170	169	187	179	153	166
Orthopedic	3	3	-	-	3	2
Ear, Nose and Throat	47	37	36	34	35	40
Pediatric	20	16	29	34	38	42
Gynecological	186	159	161	177	168	153
Other	11	11	16	8	8	5
Upper Respiratory and Heart	56	55	56	61	66	73

Table V

KAISER FOUNDATION HEALTH PLAN STUDY
HARBOR CITY HOSPITAL ADMISSIONS BY MONTH AND TYPE
JULY-DECEMBER, 1958

TYPE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER
Total	325	318	346	351	367	364
Medical	63	53	91	101	107	121
Cardiovascular and respiratory	31	24	37	55	60	64
Cardiovascular	18	16	23	36	35	40
Respiratory	13	8	14	19	25	24
Surgical	262	265	255	250	260	243

Table VI

KAISER FOUNDATION HEALTH PLAN STUDY
SUNSET HOSPITAL ADMISSIONS BY MONTH AND TYPE
JULY-DECEMBER, 1958

TYPE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER
Total	1,315	1,613	1,662	1,604	1,508	1,406
Internal Medicine	168	210	197	216	204	189
Ear, Nose and Throat	90	123	93	72	79	97
Pediatrics	273	322	384	361	341	348
Obstetrics-Gynecology	373	457	485	458	428	369
Surgery	368	430	437	392	397	344
Miscellaneous	43	71	66	105	59	59

Source: Kaiser Foundation Health Plan Records.

Table VII

KAISER FOUNDATION HEALTH PLAN STUDY
SUNSET HOSPITAL ADMISSIONS
DAILY MEANS AND VARIATION BY MONTH AND TYPE
JULY-DECEMBER, 1958

	TOTAL	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER
Total Admissions[1]							
Mean Per Day	49.9	42.4	52.0	55.4	51.7	50.3	45.4
Standard Deviation	14.2	21.4	13.6	13.6	10.4	12.8	16.0
Range	76	71	50	53	38	43	67
Internal Medicine							
Mean Per Day	6.4	5.4	6.8	6.6	7.0	6.8	6.1
Standard Deviation	2.6	2.3	2.9	2.2	2.4	2.4	3.2
Range	14	8	11	8	9	10	14
Pediatrics							
Mean Per Day	11.0	8.8	10.4	12.8	11.6	11.4	11.2
Standard Deviation	4.0	5.3	3.1	3.3	2.9	4.1	4.1
Range	22	22	12	14	11	15	16

[1] Includes internal medicine; pediatrics; ear, nose, throat; eye; dermatology; physical medicine; obstetrics-gynecology; surgery; other.

Table VIII

KAISER FOUNDATION HEALTH PLAN STUDY
WALK-IN CLINIC VISITS BY DOCTOR'S DEPARTMENT AND MONTH
JULY-DECEMBER, 1958

MONTH	TOTAL	INTERNAL MEDICINE				PEDIATRICS				OTHER[1]
		Glendale and Valley	Los Angeles	Harbor	Fontana	Glendale and Valley	Los Angeles	Harbor	Fontana	
July 7-31	17,030	1,184	3,004	1,605	2,083	1,460	1,999	340	1,721	3,634
August	22,244	1,483	4,132	2.018	2.612	1,808	2,711	442	2,332	4,706
September	20,627	1,554	3,822	2,082	2,573	1,565	2,580	300	2,121	4,030
October	23,970	1,810	4,000	2,453	3,037	2,025	2,893	287	2,588	4,877
November	22,542	1,385	3,975	2,211	2,911	1,945	2,786	317	2,643	4,369
December	24,676	1,273	4,111	2,095	3,169	2,672	2,951	407	3,640	4,358

[1] Includes principally obstetrics, gynecology and surgery.
Detailed tabulations by day are available showing individual clinic.

Table IX

KAISER FOUNDATION HEALTH PLAN STUDY
HOUSE CALLS BY ILLNESS AND MONTH
JULY-DECEMBER, 1958

ILLNESS	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER
Total House Calls	1,185	1,214	1,226	1,382	1,598	1,807
Total Respiratory	431	503	526	642	810	849
Gastro-Intestinal Complaints and Intestinal Flu	100	118	123	97	146	217
Cardiovascular	88	77	63	70	70	105
Total Respiratory and "Flu"	525	622	648	779	975	1,049
Total Respiratory and Cardiovascular	519	580	589	712	880	954

Note: Tabulations by day and detailed diagnosis are available.

Source: Kaiser Foundation Health Plan Records.

Table X

KAISER FOUNDATION HEALTH PLAN STUDY
CORRELATION MATRIX FOR OXIDANT, TEMPERATURE, CARBON MONOXIDE AND THE RATIO OF CARDIOVASCULAR-RESPIRATORY TO TOTAL HOUSE CALLS
JULY-DECEMBER, 1958

	OXIDANT	TEMPERATURE	CARBON MONOXIDE	RATIO OF CARDIOVASCULAR-RESPIRATORY TO TOTAL HOUSE CALLS
Oxidant	1.00000	0.52082	0.24218	-0.04737
Temperature	0.52082	1.00000	-0.22277	-0.22055
Carbon Monoxide	0.24218	-0.22277	1.00000	0.25973
Ratio of Cardiovascular-Respiratory to Total House Calls	-0.04737	-0.22055	0.25973	1.00000

Table XI

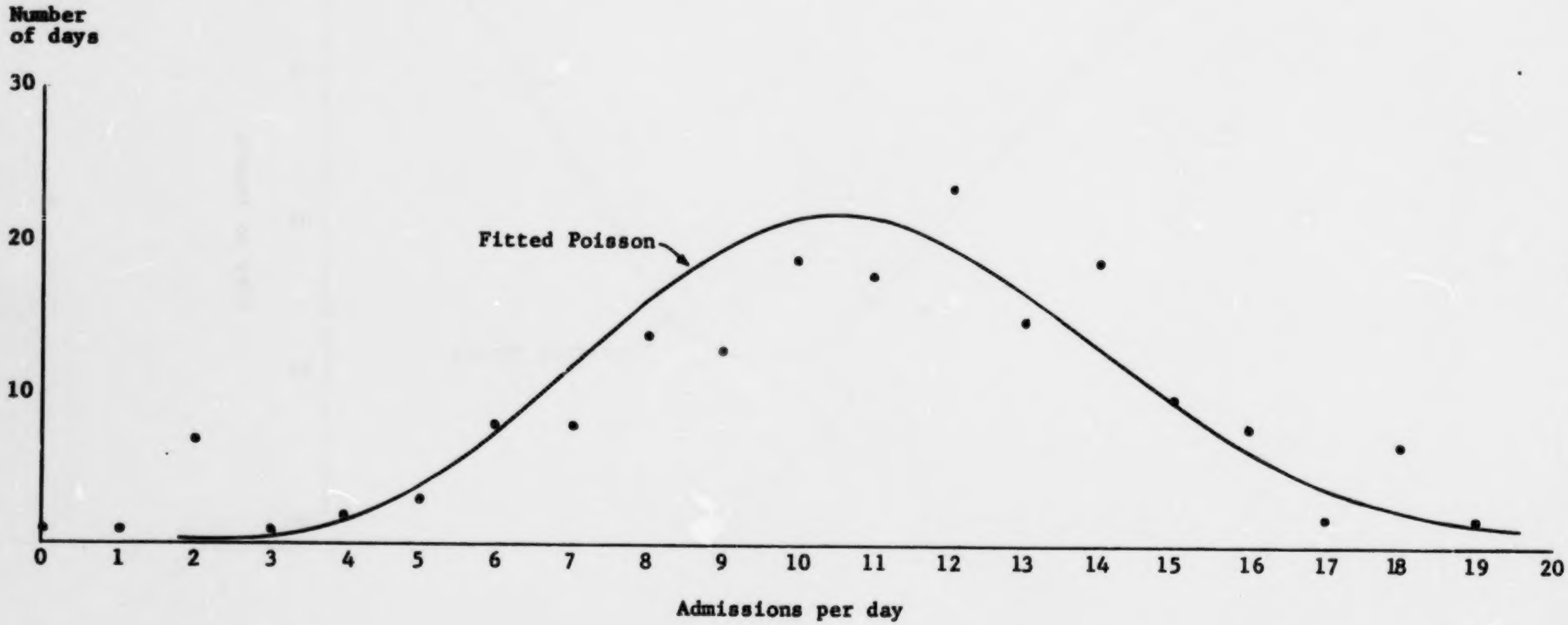
KAISER FOUNDATION HEALTH PLAN STUDY
SUMMARY OF REGRESSION ANALYSIS
JULY-DECEMBER, 1958

INDEPENDENT VARIABLES	F-VALUE	DEGREES OF FREEDOM	COEFFICIENT OF DETERMINATION	MULTIPLE CORRELATION COEFFICIENT
Oxidant, Temperature, Carbon Monoxide	6.3385	3, 180	0.0955	0.3091
Temperature, Carbon Monoxide	9.5343	2, 181	0.0953	0.3087
Oxidant, Carbon Monoxide	7.9102	2, 181	0.0804	0.2835
Oxidant, Temperature	5.2563	2, 181	0.0549	0.2343

Source: **Kaiser Foundation Health Plan Records.**
Los Angeles County Air Pollution Control District Records.

Figure VII

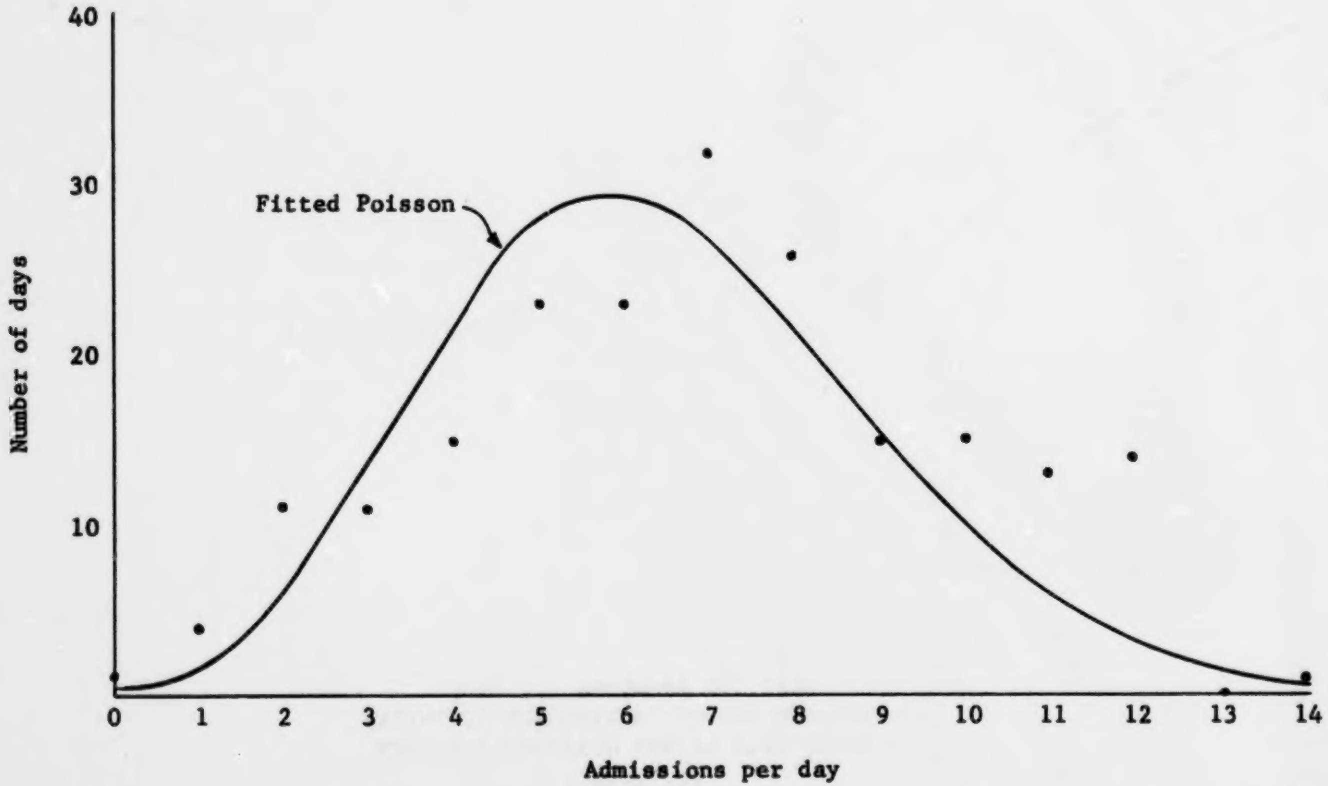
KAISER FOUNDATION HEALTH PLAN STUDY
PEDIATRIC ADMISSIONS, SUNSET HOSPITAL
July 1 - December 31, 1958



Source: Kaiser Foundation Health Plan Records.

Figure VIII

KAISER FOUNDATION HEALTH PLAN STUDY
INTERNAL MEDICINE ADMISSIONS, SUNSET HOSPITAL
July 1 - December 31, 1958



Source: Kaiser Foundation Health Plan Records.

D. Framingham Data

This Contract (No. SAph 73596) required the California State Department of Public Health to investigate the applicability of statistical methods for relating pulmonary function changes to climatologic and atmospheric pollution changes.

In attempting to meet this requirement, we became aware that it would be necessary to study the variation of pulmonary function from day to day and from person to person in order, if possible, to reduce or control these sources of variation. It was hoped that with suitable corrections for the expected variation, small differences due to air pollution exposures could be detected more readily.

Originally, over 6500 people were selected from the adult population of Framingham, Massachusetts for a long-term study of cardiovascular disease. At the onset, approximately one-third of this number failed to participate in the study. Examinations of the remainder were made every two years. By the time of the fifth examination the sample size had decreased an additional 25 percent. Reasons for this reduction in sample size were such factors as death, transfer, refusal and incapacitation.

The material and analysis reported here are based on the 3294 people who came into the clinic for the fifth examination and for whom measurements of pulmonary function were made.

History, physical and radiological examinations were made of each subject at each examination. On the basis of the examinations, individuals were then classified according to their degree of pulmonary dysfunction. Individuals who showed no signs, symptoms, x-ray evidence, or history of pulmonary involvement were classified as the "Normal Reference Population."

Analysis. For the normal reference population, multiple regression analyses were carried out for each of the pulmonary function tests taking the physical characteristics of the individuals as the independent variables. The multiple correlation coefficients for various combinations of the independent variables are shown in Table XII. It is seen that when age, height, and muscle strength are used, the addition of weight and chest diameter are of no consequence to the linear association.

The subjects in the normal reference population were grouped according to their smoking pattern and their heart disease condition. To assess the influence of smoking and heart disease on pulmonary function, the regression equation using age, height, and muscle strength as the independent variables were applied to these subjects in order to obtain their predicted values. The results were expressed as a ratio of observed to predicted and are shown in Table XIII. It is seen that smoking causes a small but significant diminution in pulmonary function among these "normal subjects".

Similarly, the regression equations were applied to the individuals in the various categories based on pulmonary signs, symptoms, or roentgenograms in order to determine the ability of the tests to discriminate between individuals with and without pulmonary conditions. The results, expressed as a ratio of observed to predicted are shown in Table XIV.

Another operation, component analysis, has been performed on some of this data. It is described in Appendix II, which constitutes a draft of an article for possible publication.

(Because these data were obtained largely through the assistance of the staff of the Framingham Heart Program, we agreed that any publication will be based upon a joint report. Therefore, we request that this present report not be quoted.)

Table XII

FRAMINGHAM STUDY
MULTIPLE CORRELATION COEFFICIENTS FOR COMBINATIONS OF INDEPENDENT VARIABLES

(Normal reference population)

	USING ALL FIVE INDEPENDENT VARIABLES	AGE, HEIGHT AND MUSCLE STRENGTH	AGE AND HEIGHT
Males			
Puffmeter	.491	.489	.463
$FEV_{1.0}$	.553	.549	.524
TVC	.646	.638	.623
1 Second Ratio	.153	.100	.096
Flow Rate	.359	.326	.326
Females			
Puffmeter	.418	.411	.407
$FEV_{1.0}$	.585	.576	.548
TVC	.719	.712	.692
1 Second Ratio	.152	.122	.095
Flow Rate	.340	.335	.311

Table XIII

FRAMINGHAM STUDY
INFLUENCE OF SMOKING AND HEART DISEASE ON PULMONARY FUNCTION

(Normal reference population)

	RATIO OF OBSERVED TO PREDICTED VALUE					
	Puffmeter		$FEV_{1.0}$		TVC	
	Male	Female	Male	Female	Male	Female
Smoking						
Nonsmokers	1.00 (64)	1.00 (274)	1.00 (36)	1.00 (147)	1.00 (94)	1.00 (466)
Former smokers	1.01 (68)	.99 (38)	1.02 (43)	.99 (20)	.98 (101)	1.00 (65)
Light	.96 (14)	1.00 (53)	.98 (9)	1.00 (29)	1.00 (24)	1.03 (102)
Moderate	.97 (114)	.93 (131)	.97 (73)	.92 (86)	.95 (175)	.98 (198)
Heavy	.95 (50)	.86 (13)	.94 (38)	.91 (10)	.96 ([illegible]2)	.92 (20)
Heart Disease						
Without	1.00 (212)	1.00 (294)	1.00 (120)	1.00 (164)	1.00 (305)	1.00 (478)
Definite	1.00 (26)	.90 (30)	.97 (16)	.96 (19)	.95 (46)	.92 (66)
Doubtful Hypertensive	.99 (101)	.97 (175)	.94 (70)	1.01 (102)	.95 (147)	.97 (288)

Note: Numbers in parentheses indicate sample size.

Source: Framingham Heart Project.

Table XIV

FRAMINGHAM STUDY
RATIO: OBSERVED TO PREDICTED

CATEGORY OF POSITIVE FINDINGS	SFX	PUFFMETER	$FEV_{1.0}$	TVC
Previous History Only	M	.96 (84)	.99 (39)	.99 (145)
	F	.98 (122)	1.00 (61)	1.01 (235)
Positive X-ray Only	M	.94 (41)	.99 (22)	.98 (77)
	F	.98 (41)	1.00 (20)	1.01 (75)
Present Symptom Only	M	.98 (72)	.93 (42)	.95 (138)
	F	.95 (103)	.95 (51)	.97 (184)
Positive Physical Examination Only	M	.88 (14)	.91 (8)	.92 (28)
	F	.93 (15)	.95 (6)	.97 (28)
History and X-ray	M	.96 (16)	1.04 (8)	.99 (41)
	F	.99 (18)	.99 (10)	.97 (38)
History and Symptom	M	.98 (21)	.98 (14)	.93 (52)
	F	.91 (47)	.90 (24)	.96 (76)
History and Physical Examination	M	.86 (9)	.93 (5)	.92 (12)
	F	.82 (16)	.85 (6)	.93 (21)
X-ray and Symptom	M	.88 (17)	.86 (9)	.94 (26)
	F	.94 (11)	.97 (6)	.95 (27)
X-ray and Physical Examination	M	.56 (2)	.24 (1)	.81 (4)
	F	1.11 (4)	1.04 (4)	1.12 (5)
Symptom and Physical Examination	M	.75 (14)	.77 (9)	.93 (30)
	F	.33 (7)	1.00 (3)	.85 (14)
History and X-ray and Symptom	M	.89 (10)	.80 (5)	.94 (21)
	F	.75 (7)	.75 (3)	.89 (15)
Physical Examination and at Least Two Others	M	.62 (40)	.65 (22)	.81 (70)
	F	.81 (15)	.72 (9)	.79 (26)
Unknown	M	.93 (111)	.90 (47)	.96 (175)
	F	.99 (121)	.97 (31)	.99 (219)

Note: Numbers in parentheses indicate sample size.

Source: Framingham Heart Project.

E. Sawtelle Veterans Domiciliary Data

The application of a discriminant analysis program to data obtained by Dr. Charles Schoettlin at the Sawtelle Veterans Domiciliary was undertaken with the results described below.

Optimum Combination of Multiple Screening Tests for Chronic Respiratory Disease. In attempting to detect and assess abnormal respiratory function, a number of laboratory procedures have been developed which may be applicable as screening tests. Unfortunately, for all of these tests, there is a rather wide range of normal values, so that an individual would have to score substantially below average results before he would be classified as abnormal. It is probable, therefore, that a person with only a mild or moderate respiratory impairment would not be classified as abnormal by a single laboratory procedure used in a screening test.

To assist in overcoming this low power to discriminate between normal and abnormal, it is possible to subject an individual to a number of different laboratory procedures and then classify him on the basis of a combination of the observed values. If the resulting discrimination were a substantial improvement over that possible using a single procedure, then the additional tests and computations might be justified.

Test values for four laboratory procedures were determined on two groups of men, one with chronic respiratory disease and one with no detectable respiratory disease. The results were evaluated using discriminant function analysis, a statistical technique which combines the test results so that the probability of misclassification for any individual is minimal. The proportions misclassified were determined for several combinations of test scores and anthropometric data and for differences between observed and expected test score values.

Data. The study group consisted of 134 men with impaired respiratory function and 147 men with no evidence of respiratory disease, selected from 2,620 men living in the domiciliary unit and chronic disease portion of the General Medical and Surgical Hospital at the Veterans Administration Los Angeles Center (often referred to as the Sawtelle Domiciliary). The 2,620 residents underwent a medical screening examination including an interview to detect selected pulmonary symptoms, a physical examination of the chest, and three trials on a pulmonary function testing apparatus, the Puffmeter. (Data used were obtained by or under the direction of Dr. Charles Schoettlin and are used with his permission.)

Men who gave histories of the presence of two or more symptoms (cough, sputum, shortness of breath, wheezing) for two years or more, and who had rales or adventitious sounds present at the time of the screening examination were considered to have chronic respiratory disease. Men with no observed symptoms and signs were considered not to have chronic respiratory disease.

Two hundred men from each group were selected using random sampling techniques and were invited to attend a clinic where, after a short series of questions and a chest examination, each man took a series of pulmonary function tests. These included three trials on the Puffmeter and two trials on the Collins Vitalometer. The maximum value for one second forced expiratory volume and for forced vital capacity were recorded, regardless of whether they occurred in the same breath.

The variables used in this analysis were age, height, weight, smoking; (of cigarettes or use no tobacco at all), maximum Puffmeter value on screening examination; and one second forced expiratory volume ($FEV_{1.0}$), forced vital capacity (FVC), and one second forced expiratory ratio ($FEV_{1.0}\%$) on the first clinic visit. The Puffmeter data from the screening examination were used in conjunction with the vital capacity measurements on the first clinic visit because in both cases it was the patient's first encounter with the instrument. This arbitrary decision introduces an excess variability associated with the fact that a few days or weeks elapsed between the screening and clinic visits.

The difference between the final sample sizes of 134 and 147 and the expected 200 in each group occurred because of men who did not attend the clinic for various reasons, men for whom information was incomplete, and men who were pipe and/or cigar smokers and/or tobacco chewers.

Statistical Analysis. The data were recorded on punched cards and subjected to discriminant analysis using the IBM 709 Computer at the Western Data Processing Center, Los Angeles.

Discriminant analysis is a statistical method used to derive a single number of Z score for each individual in two groups in such a way that the mean values for the Z score for the two groups will be maximally separated with respect to the variance within the groups. The Z score is a linear combination of a set of variables measured on each individual. Once the Z values are obtained for each individual, the proportion of individuals misclassified by using a given Z score as a dividing point between normal and abnormal can be calculated. By comparing the proportions misclassified, using different pulmonary function tests to obtain Z scores, it is possible to determine which, if any, combination of tests yields a more useful discrimination between normal and abnormal than any single test procedure.

The following sets of variables were analyzed:

1. Age, height, weight, smoking, Puffmeter value, one second forced expiratory volume, forced vital capacity, and one second forced expiratory ratio.
2. Age, height, Puffmeter, one second forced expiratory volume, forced vital capacity, and one second forced expiratory ratio.
3. Puffmeter, one second forced expiratory volume, forced vital capacity, and one second forced expiratory ratio.

In addition, prediction equations for the four pulmonary function tests based on age, height, and weight have been determined for a population sample of normal males at the Framingham Heart Disease Epidemiology Program of the National Heart Institute. Although such prediction equations for normal individuals would not be expected to be closely applicable to a hospitalized population such as the one studied, it was of interest to compare the observed and predicted values for the two groups. Therefore, the difference between observed and predicted values for each individual was subjected to discriminant analysis for the following sets of variables:

1. Puffmeter, one second forced expiratory volume, forced vital capacity, and one second vital capacity ratio.
2. Puffmeter and one second forced expiratory volume.
3. Puffmeter and forced vital capacity.
4. Puffmeter and one second forced expiratory ratio.

Results. For each of the seven sets of variables, a Z score was calculated for each of the 281 individuals. For the seven sets of variables, the proportion of individuals in the chronic respiratory disease group who would be misclassified as normal was calculated for two lower limits of normal--the score at which an equal proportion of normals and those with chronic respiratory disease would be misclassified, and the score two standard deviations from the mean score for the normal group.

The results are shown in Table XV. For comparison, similar sets of values for single tests are shown in Tables XVI and XVII.

Table XVI indicates that when the observed score for any single test is used to classify an individual, with the limits of normal referred to above, there is a large proportion of misclassification. When the difference between observed and predicted scores is used, as shown in Table XVII, there is some improvement in the proportion misclassified using $FEV_{1.0}$ and FVC, essentially no change when $FEV_{1.0}\%$ is used, and poorer discrimination when the predicted Puffmeter values is used rather than the observed value. Table XV indicates that some further improvement in discrimination can be obtained by using a discriminant function with more than one pulmonary function test.

Discussion. In the sample studied, and using the specified pulmonary function tests, it was possible to decrease the proportion of persons with chronic respiratory disease who would be misclassified as normal by a single test by using more than one test and combining the results by discriminant function analysis. The magnitude of the improvement was largely determined by the number of pulmonary function tests employed. However, even at best, the discriminant function still misclassified almost half of those with chronic respiratory disease at the limit of normal most often used in medicine. With such a high degree of error to be expected, it hardly seems justified to perform several tests and computations on each individual examined. Such a conclusion applies only to this population and may not be relevant to any other.

It is possible that some of the individuals in the chronic respiratory disease group might have had conditions which the tests employed would not be expected to detect. If the diseased group were limited to those with emphysema, perhaps a better discrimination would be possible. It is also possible that the generally debilitated condition of some of the hospitalized individuals without chronic respiratory disease may have led to low values for the "normal" group, thereby decreasing the degree of discrimination.

The technique of discriminant analysis appears to be a powerful tool for the solution of many medical problems, even though the results of this application leave something to be desired.

Summary

1. The possibility of using more than one pulmonary function test in a screening operation for chronic respiratory disease was considered.
2. Two groups of elderly men were identified as having chronic respiratory disease or not having chronic respiratory disease on the basis of medical history and physical examination.
3. A Puffmeter score, $FEV_{1.0}$, FVC, $FEV_{1.0}\%$, age, height, weight, and smoking history were determined for each individual, and several combinations of these variables were subjected to discriminant analysis.
4. Some improvement in the proportion of persons misclassified was obtained, but the degree of improvement for this population seemed insufficient to justify the multiple tests and the computations involved.

Table XV

SAWTELLE STUDY
PERCENT OF CHRONIC RESPIRATORY DISEASE GROUP WHICH WOULD BE CLASSIFIED AS NORMAL* USING Z SCORES FOR SEVEN SETS OF VARIABLES AND TWO DIFFERENT LIMITS OF NORMAL

VARIABLES INCLUDED	LIMIT TWO STANDARD DEVIATIONS FROM MEAN FOR NORMALS		LIMIT POINT AT WHICH EQUAL PROPORTION OF EACH GROUP MISCLASSIFIED	
	Percent Mis-classified	Limit (Z Score)	Percent Mis-classified	Limit (Z Score)
Age, Height, Weight, Smoking, Puffmeter, FEV$_{1.0}$, FVC, FEV$_{1.0}$%	49	-.0297	23	-.03542
Age, Height, Puffmeter, FEV$_{1.0}$, FVC, FEV$_{1.0}$%	47	-.0210	24	-.026646
Puffmeter, FEV$_{1.0}$, FVC, FEV$_{1.0}$%	52	-.0321	25	-.03720
Difference Between Observed and Predicted				
Puffmeter, FEV$_{1.0}$, FVC, FEV$_{1.0}$%	49	.01347	24	.00794
Puffmeter, FEV$_{1.0}$	56	.0147	24	.00725
Puffmeter, FVC	63	.0133	25	.00740
Puffmeter, FEV$_{1.0}$%	64	.0142	27	.00790

*The use of the word "normal" in this context is arbitrary and applies only to the operations in this set of data. The limits of socalled "normal" are not suited for any other application.

Table XVI

SAWTELLE STUDY
PERCENT OF CHRONIC RESPIRATORY DISEASE GROUP WHICH WOULD BE CLASSIFIED AS NORMAL* USING TWO DIFFERENT LIMITS OF NORMAL FOR FOUR PULMONARY FUNCTION TESTS

PULMONARY FUNCTION TEST	LIMIT TWO STANDARD DEVIATIONS FROM MEAN FOR NORMALS		LIMIT POINT AT WHICH EQUAL PROPORTION OF EACH GROUP MISCLASSIFIED	
	Percent Mis-classified	Lower Limit of Normal	Percent Mis-classified	Lower Limit of Normal
Puffmeter (liters/min.)	70	106	28	210
FEV$_{1.0}$ (liters)	67	.95	27	1.80
FVC (liters)	72	1.91	26	3.00
FEV$_{1.0}$%	80	39	31	60

*The use of the word "normal" in this context is arbitrary and applies only to the operations in this set of data. The limits of socalled "normal" are not suited for any other application.

Table XVII

SAWTELLE STUDY
PERCENT OF CHRONIC RESPIRATORY DISEASE GROUP WHICH WOULD BE CLASSIFIED AS NORMAL* USING TWO DIFFERENT LIMITS OF NORMAL FOR DIFFERENCES BETWEEN OBSERVED AND PREDICTED PULMONARY FUNCTION TEST RESULTS

PULMONARY FUNCTION TEST	LIMIT TWO STANDARD DEVIATIONS FROM MEAN DIFFERENCE FOR NORMALS		LIMIT POINT AT WHICH EQUAL PROPORTION OF EACH GROUP MISCLASSIFIED	
	Percent Misclassified	Largest Difference Considered Normal	Percent Misclassified	Largest Difference Considered Normal
Puffmeter (liters/min.)	82	$\Delta = 218.5$	26	$\Delta = 115$
$FEV_{1.0}$ (liters)	57	$\Delta = 1.594$	26	$\Delta = .88$
FVC (liters)	65	$\Delta = 1.536$	29	$\Delta = .60$
$FEV_{1.0}$%	78	$\Delta = 38$	31	$-\Delta = 15$

*The use of the word "normal" in this context is arbitrary and applies only to the operations in this set of data. The limits of "socalled normal" are not suited for any other application.

Section IV

Categories of Problems Which Have Been Recognized and Approaches to Their Solution

Introduction

There are four general problems which affect any analysis of air pollution and its effects. The first problem is that of multivariate analysis; the second, that of time-series analysis; the third, the designation of appropriate sample areas; and the fourth, the systems analysis problem. Each of these will be defined in somewhat greater detail and our experience in attacking these problems will be summarized.

A. The Multivariate Problem

It is intuitively obvious that there are a number of elements which influence the response to air pollution. Among those elements are: differential sensitivity among the population, meteorologic factors acting directly or indirectly through their effects on dilution of air pollutants, and finally, the rate of emission and reaction of the pollutants themselves. Furthermore, there are a large number of pollutants--some in gas phase and some in particulate phase. Under these circumstances, it is obvious that every detectable effect, even if it is exclusively an effect of air pollutants, is probably associated with multiple variables. The simplest possible approach is to consider all possible pairs including one environmental variable (pollutant or meteorologic factors) and one effect as though the pairs were altogether independent of each other. There is abundant evidence that this is an untenable assumption. The next simplest approach is to assume that all of the environmental variables are linearly related and not independent and that the effects are related to environment in a multiple regression scheme. Though suitable corrections for non-linear relationships can be used, the analysis soon becomes quite complicated.

In this research, we have chosen to study some multivariate problems in which there exists previous data which would suggest certain relationships. (For example, see the report of studies as to which variables were significant in determining pulmonary function in Progress Report No. 2, page 7, and in the preceding section.) One useful approach to multiple regression problems, assuming a linear relationship, is used in the computer programs Bi-Med 06, 07 and 010 which call for a replacement and deletion operation. With a replacement and deletion operation it is possible to study the effect of the inclusion or exclusion of independent variables in a linear model.

Two additional multivariate approaches of somewhat greater complexity have also been tried on data not clearly associated with air pollution. This was done because more was known a priori concerning the relationships to be expected. The first of these is the so-called discriminant function analysis. In this analysis the problem is one of combining a number of ("independent") variables with the proper weights in order to maximize the ability to discriminate between two extremes of a clearly defined and "dependent"* variable. In the example which is presented in this report, the two conditions it was desired to discriminate between were the presence or the absence of a set of signs and symptoms of respiratory disease in the Sawtelle Hospital population. In this case, the two conditions which it was desired to discriminate represented, in a sense, the two extremes of a distribution curve. On the one extreme were those persons with no signs or symptoms of chronic respiratory disease. On the other extreme were those persons with severe (at least two of the four possible symptoms and at least one sign) chronic respiratory disease. However, the analysis could apply equally well to the discrimination of any two discernible classes. For example, the analysis could be applied to days on which levels of morbidity or mortality had high or low extremes, and we would seek a linear combination of environmental measures which best discriminate between such sets of days.

---

*In the previous material, dependent and independent were used to describe the interrelationships of a set of variables. Here where the two words are used between quotation marks, the usage differs. "Independent" signifies the presumed causal importance of one or more variables with respect to a presumed resultant "dependent" on this causation. In conventional mathematical notation "independent" variables may be represented by a series of X's, i.e., X_1, $X_2 \ldots X_n$ and "dependent" ones by Y_1, as in $Y_1 = \alpha X_{11} + \beta X_{12} + \ldots \lambda X_{1n}$, etc.

In order to use discriminant function analysis, it is necessary to clearly define, a priori, the "dependent" variable.

The third multivariate method which was explored is component analysis. Here it is not essential to draw an a priori distinction between "independent" and "dependent" variables. Rather, it is desired to find several resulting variable vectors which represent the greatest amount of variation in multidimensional space, when all the available data is expressed as multidimensional vectors. Further description of component analysis will be found in the manuscript being prepared for publication. (See Appendix II). Where the number of variables is great, it is a method for reducing the number of apparent variables in a problem. An assumption of linear relationships is used; though, if one is willing to replace variables by transformations thereof, it is possible to obtain component analyses which does not require linearity.

Interpretation of fiducial limits and justification of assumptions concerning the component weights in component analysis are difficult; therefore, it is difficult to test the significance of components or weights. Further work in this area is needed. Nevertheless, the use of component analysis does seem to offer promise in future air pollution research . This usefulness would be particularly great in cases in which it is not certain whether the pollutants measured are, in fact, those emitted or reaction products derived from them.

B. Time-Series Problems

The explicit introduction of time as a variable involves the following considerations:

1. Dependence of successive observations. Classical regression analysis uses a model in which the environmental variables assume a fixed set of values and the relationships of responses to this environment are studied. This condition can probably be satisfied in laboratory studies, chamber exposure studies, or field trials, but is not usually satisfied in epidemiologic studies of air pollution. The basic assumption of multivariate correlation analysis is that each set of observations (such as the set for each of a sequence of days) of environments and responses, although possibly interrelated, is independent of every other such set. When these sets of observations follow each other in a temporal sequence, this assumption may not hold. For example, the effect of environmental variables on daily mortality in a large city may be thought to affect only a small group of individuals who are at risk, perhaps because of otherwise precarious health. Following a day of unusually high or low mortality, the number at risk may be respectively lesser or greater due to the effects of the previous day's exposure, and hence the data for the two days are not independent.

2. Variables which are functions of time. The introduction of time allows one to make use of a new class of functions of environmental variables, such as the rate of change of environment with time. Consider as an example, the possibility that the effects of temperature or pollutant levels on health may be better estimated by the rate of change of temperature or pollution than by the absolute magnitude of the variable.

3. Seasonal variation. Rather than being randomly distributed through time, most of the environmental variables follow a diurnal or seasonal cyclic pattern on which is superimposed a random type of fluctuation. Variables which vary in phase with one another will tend to seem positively correlated and vice versa. Although no causal inference may properly be

drawn from this correlation. For example, given the observable excess of deaths in winter months over summer months, positive correlations of mortality may be expected with rainfall, with levels of pollutants derived from home heating, or with levels of carbon monoxide; negative correlations may be observed with temperature, incident radiant energy, or oxidant levels. Physical, chemical, and biological plausibility are important guides to follow in interpreting this type of relationship. It may also be well to separate analytically the seasonal trend from the deviations around such a trend, and to consider the statistical operations separately for trend and deviations.

4. Delayed or latent responses. Contrary to assumptions in simple statistical operations, environmental effects on biological systems do not usually produce an immediate response, but rather a delayed or cumulative one. For example, following experimental fumigations or natural exposures, detectable lesions in plants due to air pollution often occur from the second to sixth day. It has also been shown that improvement in pulmonary functions of men with emphysema who were protected from smog exposures occurred over a period of one to two or more days.

The primary emphasis in time-series studies has been on the items outlined in 1, 3 and 4 above. Work on time-series methods will be carried forward with the assistance of a research grant recently received. Our principal attack on this problem has been on the third item above, the problem of cyclical regression, for which a specifically prepared program was developed under this contract. (A copy of the program is appended--Appendix III.)

C. Air Pollution Area Coding Problems

When one wishes to study the possible association of air pollution exposures with some response of human or other subjects, the question always arises as to what the exposure of the subjects might have been. Since a relatively small number of air pollution monitoring stations are distributed within the geographic confines of a large geographic basin, one attempts to relate the measurements at a given station to a certain group of individuals who work or reside in the air volume sampled by the station. Furthermore, most of the monitoring stations are located in the open air while the population spends part of the time indoors. Further complicating this problem are such matters as whether the maximum, median, or time-weighted average value should be used, or whether a threshold exists.

From a practical point of view, it is important to define reasonably homogeneous areas within a basin and to determine which sampling stations represent the air quality associated with such areas (See also p.10 above--nursing home). Hills or mountains would represent a source of non-homogeneity to an area. The surface contrast of high buildings, low houses, grass land, and forest land also represent a source of non-homogeneity. So, too, may river valleys, or areas where industrial activity contributes an important part of the pollution. The experience in Los Angeles with trajectories of parcels of polluted air suggests that pollution in one part of the basin may readily affect persons in geographically distant areas of the basin or even of adjacent basins such as those in the San Diego area. In general, the sampling sites relevant to the exposure in a given area should be upwind of that area. However, it is not easy to define "upwind" when gentle winds exist which shift to many points of the compass during the twenty-four hours. Furthermore, air pollution concentrations are greatest during just such conditions. The selection of homogeneous areas and representative sampling stations should be based on the analysis of available data, such as is being done in the "Two-Community Study."

The following section describes some specific techniques which have been used to delineate areas for comparison of morbidity and mortality.

Coding by air pollution areas. In order to compare measures of morbidity or mortality among populations experiencing different exposure to air pollution, it is necessary to categorize these populations in terms of exposure to air pollution. This is usually accomplished by dividing the area being studied into smaller areas, each being relatively homogeneous with respect to air pollution, but differing from each other.

Criteria for delineating geographic air pollution areas. Some criteria for delineating air pollution areas are listed below:

1. Satisfactory air pollution measurements should be available for each area for the time period under study.

2. Each area should be relatively homogeneous with respect to air pollution for the time period under study.

3. Other variables should be controlled if possible; that is, areas should be delineated so they are relatively homogeneous with respect to other environmental variables (such as temperature) and social variables for the time period under study. In any event, these variables must be considered in the analysis.

4. The size of the areas should be appropriate in terms of expected numbers of deaths or cases of disease.

5. The areas should permit relatively simple coding by address or by other locational data (e.g., census tract, telephone exchange).

The emphasis on time in points one, two and three is a caution against using air pollution areas which have been used previously or for which data are conveniently available but which may be irrelevant for the time period being studied. Irrelevancy of areas previously established may also occur with respect to points four and five; that is, size of areas with respect to expected numbers, and ease of geographic coding.

Census tracts and the Two-Community Study. If a mechanical method is available for coding to census tract or if data have been already coded to census tract, the larger air pollution areas may be formed by the combination of individual census tracts. The use of census tracts is convenient if social characteristics are to be used in the analysis, although areas defined in other terms could be characterized by the characteristics of the census tracts included even though the boundaries do not necessarily coincide with tract boundaries. Census tracts were used as the primary unit in the determination of areas used in the Los Angeles County Two Community Study. The final air pollution map used in this study, however, was drawn in terms of groups of census tracts (statistical areas) rather than individual tracts. Air pollution measurements, plant damage, and complaints of air pollution were used to designate the areas experiencing high, moderate, and low air pollution. These areas were used in the nursing home mortality and hospital transfer study although they are based on 1956 data and are not applicable for the entire period of the nursing home study.

Hospital Admissions Study. For use in the study of hospital admissions throughout Los Angeles County, the County was divided into the following five areas: Central Los Angeles, West Los Angeles, Southeast, Southcoast, San Gabriel Valley, and San Fernando Valley. Similar areas, although not precisely the same, were used as a second method of coding the nursing home data. Some of the areas do not contain a large enough proportion of the total nursing home beds in the study to provide useful results.

Areas containing clusters of nursing homes. The two geographic coding schemes described above are especially useful for studies in which the entire population is exposed to risk; that is, may die or be admitted to a hospital. On the other hand, in the nursing home study, by definition, only a part of the population is exposed to the specific risk being studied. Therefore, it seems logical to define the air pollution areas in terms of the geographic clustering of the homes. This is the third method used in coding the nursing home data, and it resulted in the following areas: Central Los Angeles, peripheral Los Angeles, Santa Monica, Long Beach, Pasadena, Temple City, and a residual area including widely scattered homes. The central Los Angeles area includes between one-third and one-fourth of the total number of beds in the study. If a circle is drawn including these homes, two air monitoring stations are located close to the circumference on approximately opposite sides of the circle. The peripheral area surrounds the central area and should probably be used in combination with it to form an expanded central area. The other areas contain fewer nursing home beds. Unfortunately, the residual area which is difficult to define in terms of air pollution exposure, contains about as many beds as the central Los Angeles area.

Interpretation of Area Differences in Mortality. Tables XVIII and XIX are included as an example of differences in average weekly mortality in nursing homes in different geographic areas. If these areas truly do experience significant differences in air pollution and if other variables are ignored, one might conclude that the sea breezes of Santa Monica increase the chances of mortality among occupants of nursing homes, while the smoggy downtown area is actually beneficial to health. But the statistician is not allowed to neglect the other variables, and reflection calls to mind such factors as socio-economic selection of occupants, admission policies with respect to degree and type of disability, policies of transfer to hospitals, and age of occupants. The next step is to devise ways of taking into account these variables and determining whether there remains a residual difference in mortality which might logically be attributed to air pollution.

It should be pointed out that Tables XVIII and XIX give no indication of what may be occurring within each year. It would indeed be convenient to find that differences among areas increased during periods of heavy air pollution in one or more areas. It might then be assumed that the other variables did not change significantly over time within each area in such a way as to mimic this relationship to air pollution. This is the approach used in the Two-Community Study.

Table XVIII

LOS ANGELES COUNTY NURSING HOME STUDY
DEATHS BY NURSING HOME CLUSTER AREAS
JULY, 1955-DECEMBER, 1959

	TOTAL	CENTRAL LOS ANGELES	PERIPHERAL LOS ANGELES	SANTA MONICA	LONG BEACH	PASADENA	TEMPLE CITY	RESIDUAL
1955 (July-December)	1,126	319	168	90	75	103	49	322
1956	2,095	532	284	174	137	217	120	631
1957	2,101	568	251	138	148	187	160	649
1958	2,744	574	304	212	263	241	263	887
1959	2,562	540	299	245	191	224	210	853

Table XIX

LOS ANGELES COUNTY NURSING HOME STUDY
AVERAGE WEEKLY DEATH RATES BY NURSING HOME CLUSTER AREAS
JULY, 1955-DECEMBER, 1959

	TOTAL	CENTRAL LOS ANGELES	PERIPHERAL LOS ANGELES	SANTA MONICA	LONG BEACH	PASADENA	TEMPLE CITY	RESIDUAL
1955 (July-December)	11.5	10.1	10.9	21.6	16.4	16.0	5.6	12.1
1956	10.7	8.4	9.3	20.8	14.2	16.4	6.9	12.0
1957	10.4	9.0	7.9	16.5	15.3	14.0	7.2	12.2
1958	11.5	8.9	8.8	13.9	17.7	13.1	10.7	13.1
1959	10.4	8.6	8.8	15.3	13.0	12.4	7.2	12.0

Note: Rates are per 1,000 nursing home beds.

Source: Nursing Home Records.

D. Systems Problems

It is obvious that the selection of any arbitrary number of variables analyzed in any conceivably elegant fashion may tend to be an inadequate explanation or description of the complex problems associated with air pollution. For example, there will also be variations in traffic patterns, there will be variations in motor vehicle operating conditions associated with weather, there will be variations in fuel consumption associated with weather, there will be variation in the rapidity and intensity of photochemical reactions associated with alterations in the angle of incidence of sunlight, there will be variations according to the vertical mixing velocity of the air, there will be variations associated with the presence of a variety of natural and photochemical aerosols, etc. All of these elements, measurable and unmeasurable, may be thought of as the "system" related to air pollution. Systems analysis is a method for dealing with these complexities and uncertainties.

The tremendous number of elements which contribute to variations in air pollution and its effects suggests that analysis of the most crucial factors which influence this system should be given preference. The relatively newly developed methods of systems analysis have been suggested for application to this problem. Work under this research contract has not been undertaken in this area, but we feel that serious efforts should be undertaken to develop an accurate model of atmospheric emissions, reactions, and effects which include all of the potentially measurable processes and constituents. Under these circumstances, as data are accumulated concerning the relationship of one or more variables with each other, it would become possible to come closer and closer to understanding the interaction of a fairly large number of elements and variables. Such an approach will also permit a selection, upon a rational basis, of important areas for further investigation. This can be done even when a large number of the relationships between the elements are not known. It is therefore recommended that further attention be given to systems analysis in ongoing statistical research upon the health effects of air pollution.

Section V

Statistical Methods in Air Pollution Medical Studies
A Review and Appraisal

Statistical Methods in Air Pollution Medical Studies

A Review and Appraisal

Introduction

Man-made air pollution has long been recognized as disagreeable esthetically and possibly harmful to health. Specific industrial contaminants (silicates, fluorides, soot, etc.) were recognized early as injurious to health when occurring in high concentrations in limited airspaces such as mines, foundries, etc. It was not until the Meuse Valley (1930), the Donora (1948) and the London (1952) episodes that air pollution in the ambient atmosphere of large urban areas reached levels sufficient to cause health effects (discomfort, morbidity, and mortality) detectable by the routine methods of epidemiologic surveillance of standard morbidity and mortality reports. Since these episodes there has been no serious doubt that it is possible for a city to produce concentrations of air pollutants that are severely irritating, especially to mucosal surfaces in a large proportion of the resident population and that result in morbidity ranging up to and including premature death in certain especially susceptible or fragile subsets of a population.

Since then, the problem of research in health effects of air pollution has been one primarily of specification and titration. It is important both to our scientific understanding of health effects of air pollution and to our administrative procecures for regulating it to discover:

a. what specific contaminants cause what specific health effects,
b. in what specific categories of a population these health effects occur,
c. what the thresholds and dose response relationships are for each such contaminant and combination thereof, and
d. what low levels of contamination may be regarded as harmless if sustained or recurrent over a long period; i.e., the long-term cumulative dose-response relationship at low levels.

Epidemiologic research in recent years has thus concentrated on efforts to detect: 1) immediate, 2) delayed, or 3) cumulative effects of low level atmospheric pollution through exploitation of extant morbidity and mortality data on the response side, and urban air pollutant measurements and meteorologic data on the dose side.

The U.S. Public Health Service report by Mancuso and Mordell outlined in great detail the possible data sources, their limitations, and broad approaches to the epidemiologic investigation of extant data. Many of these suggestions have been pursued. The basic approach suggested is the correlation study in which one observes variations in levels or amounts and kinds of air pollution and seeks concomitant variations in various available health indicators. Two general methods are available, which might be called 1) the geographic method and 2) the time-series method.

> 1. The geographic method looks for relationships between measurements of air pollution and of health as these vary together across contrasting geographic areas (or political subdivisions). The geographic areas may be areas of high and low pollutant concentration within the same city, or different cities across the country. Recently, Landau has outlined still further methods for nationwide geographic epidemiology in air pollution studies using existing demographic and mortality data.
>
> To date, this approach has not yielded more than already obvious results, possibly because of the difficulty of standardizing the reporting of air pollution data from enough cities to make inter-city comparisons sufficiently sensitive. The contrasting areas-within-city approach has been tried in Cincinnati, Nashville, Los Angeles and elsewhere with negative or inconclusive results, apparently because of difficulties in obtaining adequate health and air pollution data on the same sub-areas within a city, or of getting sub-areas that are small enough to be homogeneous and truly distinctive and yet large enough to permit adjustment and standardization for the many co-variables (age, sex, race, occupation, etc.) involved.

2. The time series method follows health and air pollution indicators for a defined geographic area or for a defined population over time, usually but not necessarily at equal intervals (hour, day, week). Here the population at risk is assumed to be relatively constant over time, the air pollution dose varies from hour to hour and day to day, and immediate, delayed, or cumulative effects on health indicators are observed as a co-variable with the air pollution indicators. This approach has been pursued energetically, still with inconclusive results so far as low level effects are concerned.

The analytical methods used (classical correlation and auto-correlation) are appropriate and are adequate for detecting quite strong relationships or large effects. Their failure to do so reduces the problem to one of discovering rather small effects in the presence of large random variability, and the question arises whether more powerful concepts of experiment design and statistical analysis might be fruitful.

The purpose of this paper is to discuss this question in the light of the state of our knowledge of air pollution and health measurement as well as the state of the air of statistical analysis.

Structure of the Problem

The problem may be thought of analytically as essentially a bioassay in which human subjects are exposed to various doses of toxic substances and their responses recorded so as to determine the functional relationship of response to dose, $R = f(D) + e$. But in the air pollution epidemiological studies the situation is, in several important respects, quite different from the neat arrangements that can so often be achieved in a laboratory dose-response experiment. For example:

a. The dose is a random variable, fluctuating over time and space and not subject to experimental setting, as in laboratory determinations.

b. The dose does not vary over any very great range of effectiveness. In the Donora crisis, the dose was sufficient to produce chest constriction in as many as 40 percent of the population in the age groups over 55. The highest mortality observed in any group was 5.7 percent (5 deaths out of a population of 87 in the non-White population in the 55-64 age group). In Los Angeles during the severest peaks of pollution, something like 90 percent of the population may be affected in certain sub-areas, with respect to eye irritation, but no detectable level has been reached with respect to mortality, even in presumably highly susceptible population subsets. As a result of this limitation, it may be inherently impossible to estimate the slope of the dose response curve, or even to detect any response at all, except in the case of presumed minor discomforts such as eye irritation and visibility.

c. The dose and the response can both involve several measurements, some of which are known to be interrelated, such as ozone concentration and temperature. Still other unrecognized relationships must be presumed to exist. Further, both sets of measurements are multivariate in form rather than being of the comparatively more tractable multiple univariate form in which some of the variables are fixed or selected by the experimenter. In the multivariate situation, the several measurements vary jointly in some random fashion not subject to control by the experimenter. Although serious difficulties arise in the analysis and interpretation of multivariate observations, this kind of "experiment" does have an offsetting advantage in that measurements are taken at levels and in combinations that actually occur.

To illustrate the differences between multiple univariate and multivariate experiments, suppose we wish to study the functional relationship of mortality (D) to temperature (T) and ozone concentration (O). We might think of this as either kind of situation, depending on how the data are collected and on how the analysis is to be interpreted. We might, for example, select or fix values of T and O separately in a factorial arrangement with T set at several levels, each of which is crossed with several levels of O. A response D could then be measured at each of the T and O level combinations. Here T and O appear as two separate, not necessarily independent or additive variables, both selected by the investigator. The usual multiple regression analysis (to be reviewed later) takes a model (e.g., $D = aT + bO + e$) involving parameters a and b and random, residual effects, e. The variation in the random effects reflects errors of observation, the usual variability of the thing measured, or effects of variables left out of the model. This, in probability or stochastic sense is really a univariate situation in that only one of the variables on the right side, e, is subject to chance variation. The whole expression aT + bO + e is therefore called a single random variable, or a variate. There are, however, three variables in the mathematical description and thus it is sometimes called multivariable, meaning that several variables are involved. A multiple regression analysis would be appropriate.

Alternatively, and more typically in air pollution health studies, we might have observed sets of concomitant values of D, T, and O as they chanced to occur in the community. Here the three values are jointly subject to chance variation and we have a multivariate situation. We might still be interested primarily in a functional relationship such as D = aT + bO + e and, if so, most of the analysis would be done as in the multiple regression model. Another point of interest might be the relationships between the two environment variables (temperature and ozone) in this case, particularly if the number of such variables were large.

In environmental air pollution health studies the dose consists of numerous constituents in complex mixtures, many of which cannot be varied separately, and the response in the human subject may also require several different measurements for description. It is by no means clear that a single dose parameter, such as air visibility can adequately integrate the many variables that are operative. They could conceivably cancel or mask one another. Hence, we have to deal explicitly with a multivariate dose and probably with a multivariate response.

Complications

A. The dose. The dose, $R = f(D) + e$, may be thought of as the blanket of air, up to a few hundred feet above ground level, surrounding the host, be it an individual person or a whole city population. Composition of this air is complex and continuously changing in space and time. For our purposes, two main components may be identified: "pure" air and pollutants.

1. "Pure" air constitutes most of the dose by volume, acting as a neutral vehicle and diluent of the pollutants, as the denominator for expressions of concentration of the pollutant dose.

2. Pollutants.

a. Physical ingredients. These may be large particles, which settle out rapidly and are of importance only near the emission source, or they may be small particles, which become suspended in the air or form aerosols and hence are available for chemical reactions in the upper air and also may impinge directly on tissues of the human host. These physical particles include pollens and dusts from "natural" sources, as well as man-made particles. Some particles have been demonstrated to be small enough to pass into the smallest alveoli in the human lung.

b. Chemical ingredients, gases. The usually measured and monitored chemical ingredients are carbon monoxide, nitrogen dioxide, total oxidant, ozone, and sulphur dioxide. In addition to these there are known to be various olefins from natural sources, olefins from petroleum combustion, aldehydes, traces of fluorides and heavy metals such as lead, zinc, etc.

These pollutants, or potential toxicants, may be considered singly in relation to health effects in the host, but such an approach appears to make sense only in rare and usually quite local episodes of severe and preponderant contamination by a single toxicant. In the more usual situation of day-to-day low level pollution observed in our cities these ingredients are in a highly dynamic system of physical and chemical interactions and mixtures as well as to the amounts of separate ingredients. Hence provision should be made for treating the dose as a multivariate system, in which the observations are samples out of the joint distribution of several variables.

3. Meteorologic conditions. The characteristics of the dose are highly influenced by the whole system of weather variables, which also interact among themselves and with the pollutants in a complex, dynamic way. The principal weather variables which are usually recorded routinely are:

Air temperature: daily maximum, minimum, average, deviation from historic norm, degree days.

Air temperature gradient with altitude: temperature inversion--presence or absence, altitude.

Humidity: daily maximum, minimum, average.

Precipitation: daily total amount.

Air pressure: daily maximum, minimum, average.

Air movement: daily wind--fastest velocity and direction, fastest mile.

Sunshine: daily total and percent of total (hours) possible, percent sky cover.

These measurements, though seemingly fairly direct and straightforward, may in fact be quite indirect indicators of the real operating variables. Temperature is taken near ground level, for instance, which may be pertinent to direct host response but only a very indirect measure of the higher level temperatures where the pertinent chemical reactions are taking place to produce the dose. The crude observation of sunshine is known to be a poor indicator of incident ultraviolet radiation because of absorption and scattering by suspended dusts and particles that takes place at all levels, including very high levels imperceptible to the usual meteorologic observations.

The dose side of the equation is further complicated by two considerations that present nearly insurmountable obstacles to experimental determination of the dose-response relationship:

a. Incomplete mixing of the air means that we can have no assurance that the air that is observed at monitoring stations is the same as that to which the host is actually exposed.

b. Confounding activity of some aerometric factors is serious. While temperature and humidity conditions are important determinants of the air pollution chemical dose, they also are biologically active in the human host through mechanisms unrelated to air pollution. Extremes of temperature and humidity are known to cause excess morbidity and mortality in their own right. Hence these two patterns of biological activity of air temperature and humidity are inextricably confounded in epidemiologic investigations of the air pollution dose-response relationship. Some early air pollution mortality studies naively attributed to the concomitant air pollution excess mortality or morbidity which could, so far as the statistical evidence went, probably be accounted for by the temperature fluctuations alone.

4. The indicators of dose. What can be measured as indicators of the dose is the amount of each of the pollutants per volume of air (concentration) or per unit time (amount) observed at each of several monitoring stations, within a city, or in different cities.

The following characteristics pertain to the dose:

a. These variables fluctuate in time, subject to diurnal, weekly and seasonal cyclic influences and to secular trends.

b. The fluctuations in dose may be subject to a certain amount of feedback damping,in the sense that when some arbitrarily set levels are reached at monitoring stations an "alert" is declared. The "alert" could be followed by a reduction of the level, though with a variable time lag. This deprives the investigator of the most informative data points for his assay; since presumably the higher levels of the pollutants cause the most marked responses.

c. There is auto-correlation among values observed at successive times, so that observations of successive hours or days may not be considered independent. Auto-regression and lagged-correlation studies (Hechter) have indicated that the correlation fades out rapidly after the first day or two, but is not negligible for adjacent days or hours.

d. The pertinent dose might be:

1. the instantaneous level (concentration)
2. the total cumulative amount, averaged in some sense over a short period, such as a few hours or days, or over a long period, such as several years,
3. complex simultaneous combinations of pollutants, or sequences of particular pollutant doses over time. Study methods need to be sensitive to a wide variety of patterns of the dose.

e. The dose leads the response by an unknown lag period that might vary from a few hours in the case of acute responses to instantaneous or short cumulative doses, to several years in the case of responses to long accumulations.

Methods of analysis should so far as possible take into account these many pecularities of the dose parameters.

The Response

1. The unit host may be considered almost any biological level of the organism: a single epithelial cell, an alveolus, a bronchus, a lung, a trachea, a conjunctiva, a whole person, or a defined population or population subset of persons. As in any sampling situation, experimental reliability may be expected to be somewhat proportional to the homogeneity of the host units, so that experiments using cells and organs as the host units might be expected to yield more reliable (more sensitive) results than experiments using whole persons or sets of persons as the host.

2. Host heterogeneity having an effect on response variability is present at all levels, but a number of assignable sources can usually be thought of, such as: genetic constitution, physical factors such as size of organ or person or population, sex, age, race, occupation, socio-economic situation, habits, activities, health status, etc.

3. Response indicators. Irritation and/or irreversible damage to the host unit may potentially be observed at each of the levels of unit host. A considerable range of response indicators is available for study. It is important to note that an indicator such as plant damage or irritation of conjunctivae may be a direct end-point measurement if the plant or the conjunctiva is considered the unit host, but a quite indirect measurement if a whole person, or the population of an area is considered the unit host. These response indicators cannot, therefore, be labeled *a priori* direct or indirect. Response indicators that have been studied or appear to be accessible are:

a. Mortality indicators
   - Number of deaths
   - Crude mortality rate for geographic or other subdivision
   - "Excess" mortality over "normal" expectation
   - Cause-specific mortality rates
   - Age-specific mortality rates

b. Acute sickness
   - Asthma attacks in asthma patients
   - Bronchitis attacks
   - Exacerbation of emphysema patients and other pulmonary cripples

c. Presumptive illness
   - Other cause-specific morbidity rates
   - Clinic admissions
   - Clinic walk-in visits, especially such category clinics as eye, throat, chest, cardiovascular
   - Hospital admissions
   - Nursing home deaths or transfers to hospitals
   - Absenteeism from school or work
   - Accident rates--home, vehicle, or industrial

d. Untoward symptoms (which in the absence of an apparent cause might lead a person to seek medical attention)
   - Cough
   - Eye irritation
   - Chest tightness

e. Insidious or chronic disease
   - Community morbidity rates for lung cancer, bronchitis, emphysema, fibrosis, etc., for long-term residents

f. Alteration of important physiological functions

Changes in measured pulmonary function
Changes in visual dark adaptation

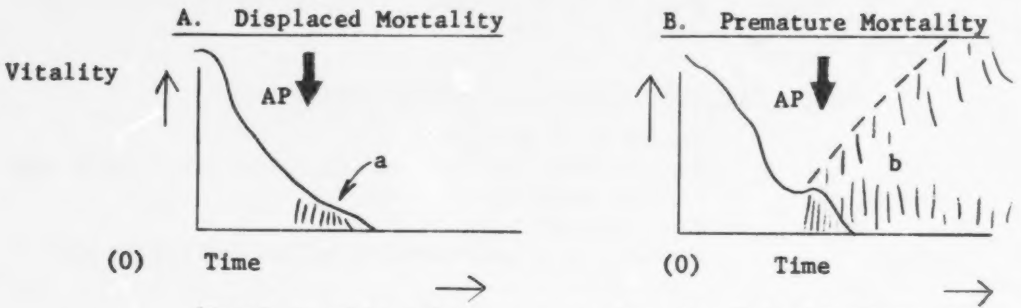
g. Discomfort sufficient to lead to change of residence or employment

Rate of such changes, as measured by California Health Survey

h. Special experimental responses, such as

Tissue culture damage
Conjunctival irritation, experimental animals and human beings
Plant damage
Domestic and experimental animal morbidity and mortality

The most definite serious health effects of air pollution so far documented have been derived from mortality data. Skepticism is frequently voiced about the importance of a slight excess of mortality, even though statistically significant, as a measure of pollution damage. If such an excess, for instance, occurs only in the aged and infirm, and an excess period is followed by a compensating period of deficit mortality, it is argued that an environmental agent that merely hastens already inevitable death by a few days is trivial as a public health concern. Be this as it may, there is another kind of air pollution mortality that could be quite important in social cost. This "displaced mortality" might be represented by a curve A, illustrating a person on an irreversible terminal trend towards death, with the air pollution burden simply hastening the end. It is perhaps not altogether proper to charge such deaths to air pollution. At best, only the small remaining-life segment (a) should be charged to air pollution. On the other hand, a heavy air pollution episode may cause some deaths of an entirely different kind, which we might call "premature mortality", represented by Curve B. This is the new-born, the young or middle-aged person critically ill from disease, trauma or operation, who is about or beginning to recover, and eventually will be restored to full vitality, with a substantial remaining-life in store. During this critical period an air pollution effective dose might cause death, resulting in a really substantial loss of future life, (b).


A. Displaced Mortality

B. Premature Mortality

(Death in this scheme occurs when the "vitality" reaches zero.)

In efforts to assess the social cost of air pollution health effects it will be necessary to distinguish these two kinds of death. A single type B death might cost more man-years than a large number of type A deaths. Wherever daily or weekly excesses can be detected, as at Donora, New York, London, analysis of this question should be attempted.

Dealing with the Complications

The many complications have the effect of concealing the relationship we seek to determine amidst a whole confusion of nuisance variables, or "noise". Methods of singling out the specific relationship from this confusion require some form of dose amplification, response amplification, or filtering out the extraneous noise.

a) Dose amplification. Selection of environment areas with especially high pollution levels is a dose amplifying move. Accumulation of data on panels or cohorts of long-term residents, though slow and expensive, has the effect of amplifying the totel amount of the dose. It may, however, dilute the instantaneous effect by averaging over periods of low as well as high dose.

Creation of an artificially controlled environment, such as a closed weather box or room, with intensified dose levels is an approach of the experimental investigator, It is limited, however, to study of particular pollutants one or a few at a time. Urban air is so complex and so dynamic a function of so many pollutant, aerologic, meteorologic, and topographic factors that the likelihood seems small of producing literal physical simulation of more than a very simplified model of the real situation. When human subjects are involved there is the additional limitation of legal restraints on subjecting them to more than usual doses of potential toxicants.

Improved dose measurement could help. The pollutant measurement made by air pollution monitoring stations may be quite indirect as regards measurement of the dose to which the host is actually exposed. Ideally we would like to have a direct measurement for each member of the host population, similar to the dosimeter patches used by workers in radiation and X-ray laboratories.

b) Response amplification. Selection of especially sensitive population subsets, such as premature infants, pulmonary and cardiac cripples, etc., may raise the response frequency to detectable levels, even with the rather low dose levels provided by expected environmental air pollution.

More sensitive and precise measurement of responses might be found in some instances. Advances are being made in the instrumentation of pulmonary physiology and these may result in some response measures somewhat more precise than the present measures, which have large errors of determination.

Search for additional patterns of responses could conceivably find some that are more specifically sensitive to air pollution than those now in use.

c) Filtering out noise. Adjustment for the effect of extraneous variables can be in various ways:

1. Epidemiologic controls for contrasting populations or population subsets can be sought which are similar on some of the nuisance variables but different with regard to exposure to the main air pollution variables under study. This two-community type of study plan, though straightforward in concept, has turned out to be difficult in practice. It is nearly impossible to find pairs of communities that are alike in almost every way except their air pollution patterns. If such a community-pair can be found, analysis would then be a multiple regression of the inter-community difference in daily (or weekly) deaths on differences in temperature and air pollution indicators. Professor Savage, who proposed numerous refinements of this model, showed that it was probably impracticable. Since the excess mortality due to low-level air pollution is most likely rather small, it would

take very large communities--city-size--to make such a study plan sufficiently sensitive. If different cities are paired, instead of different segments of a single city, further confounding of intercity differences is introduced.

2. Experimental controls. Some parts of the problem can be isolated and investigated experimentally using methods, more typically found in the laboratory, of excluding or controlling unwanted variables. In some limited ways, to be mentioned later, experimental controls can be imposed in the field situation as well as in the laboratory, and may offer better yield for the effort than the available epidemiologic controls.

3. Statistical Analysis for control. When the unwanted associated variables cannot be physically eliminated or held constant, they still may, if measured and recorded, be "controlled" or adjusted for in the statistical analysis. It must be emphasized, however, that the power of statistical analysis to impose the necessary controls and separate out the variables desired depends on the way the data are collected and on how accurately the adjustment model represents the physical situation. It is never certain that some other variable not considered in the adjustment is the important one.

Usually, there is not much that can be done with data collected as available. Elaborate *ex post facto* schemes for adjusting and standardizing data impose a very heavy burden on the logic of the statistical analysis and on the credulity of the scientist. In order properly to exploit the adjustment and filtering possibilities of statistical analysis in epidemiologic investigations, it is essential to plan and "design" the observation and data collection just as carefully as for an artificial or laboratory experiment.

5. Quality of the data. The observations and measurements used in air pollution health studies vary widely as to quality. In the past much use has been made of mortality, morbidity and weather data originally collected for other purposes. As we move into the study of present and future air pollution, new and more specific data sources may have to be developed. With such needs in mind, the current sources are reviewed.

Routine Mortality Data.

The completeness, availability, and quality of vital statistics obtained from death certificates is generally high for the main data such as age, sex, place of residence and place of occurrence. Cause of death is probably sufficiently reliable for the broad classifications, particularly the cardio-vascular-respiratory group of such interest in air pollution studies, but when individual deaths are studied, much more detailed and accurate information is required. For place of occurrence and of residence the death registration data are good for the larger political subdivisions but of less reliability as to within-city designations, such as census tracts. One highly desired variable, length of residence, is seldom obtained.

Despite some shortcomings, official mortality data will continue to be a major resource in air pollution studies. Their greatest use will no doubt be in very large geographical epidemiological and long-term time-series studies. Because of the small numbers of daily deaths in even large cities, they will be of limited use for within-city studies.

Routine Morbidity Data

The reporting of "reportable" morbidity to local health departments is generally so incomplete as to provide little information of use in air pollution research.

Admissions to hospitals, nursing homes, and clinics have given some promise as an indirect indication of acute effects of air pollution. Yet, even in a large city like Los Angeles the numbers in disease categories specifically pertinent to air pollution are rather small, with the result that not much adjusting or analysis can be done to extract the rather weak air pollution effects from among the many associated or nuisance effects. Pilot studies of daily hospital admissions and walk-in visits at large out-patient clinics using the routine institutional records failed to detect any association with air pollution indicators. However, it may be that too short a time period was studied for elimination of trends, cycles and perturbations caused by extraneous factors. Hospital and clinic records kept for administrative purposes seldom are sufficiently complete and accurate for research purposes.

Industrial and school absenteeism rates, like hospital and clinic admissions, might reflect air pollution effects, but they are also subject to many other influences. Industrial and traffic accident reports are available but suffer from the same limitations.

National Health Survey morbidity data may eventually be available in sufficient detail and nationwide uniformity to be useful for inter-city geographic epidemiology studies, but because of their rather sparse sampling in time (as well as in space) it will be of little help for within-city studies.

The California Health Survey, begun in 1955, measures, among other things, the attitudes of residents towards air pollution and subjective feelings of effects of air pollution on well-being. These measurements, based on questionnaires and interviews, have been subject to all the well-known difficulties of opinion surveying, but they have been able nevertheless to discriminate geographic differences fairly effectively. Information obtained by the Survey on changes of residence attributed to air pollution provides one of the few definite effects of low-level chronic pollution so far detected by population survey methods. It is possible to argue whether this represents a real health effect.

In sum, use of extant mortality and morbidity data, even when the data quality is inherently good, has been disappointing with respect to detection of health effects of continued low-level pollution. An appended bibliography* exhibits a wide range of efforts to exploit such data, with almost unanimous failure to extract health effects due to air pollution from among the many other sources of variation governing such data.

Special Survey Data.

Special morbidity surveys have been attempted in an effort to obtain a highly susceptible subset population, such as asthma or emphysema patients. Unless extreme care is taken, data from such studies will usually include much of the extraneous variation encountered in the more general morbidity data. In longitudinal studies the added difficulties of panel retention and follow-up need to be recognized. Because of the great expense of making direct daily observations of patient status, instrumentation is sometimes resorted to, such as patient diary, telephone or postcard questionnaire. Such instrumentalities inevitably introduce substantial inaccuracy into the data, not all of which can be removed by occasional direct observation.

Controlled Experiment Studies.

Quantitative physiological and biochemical studies of individual persons (normals or especially susceptible cardio-respiratory cripples) exposed to environmental air pollution uniformly encounter very large variances in their measurements, even within the same person from day to day, that presumably are not related to air pollution. Instrumentation in these studies has encountered difficulties, it being difficult to devise a system of measurement of respiration functions that is independent of the subject's voluntary control. The standard values (norms) in current use for clinical work usually turn out to have been determined on too small or too special a sample, or are too imprecise for use as guides in experimental work. Hence, each investigator is under the practical necessity of doing much exploratory and calibration work to establish norms for measuring systems before they are ready for application to the study of pathological responses.

---

*The bibliography is not included but will be appended to this material when it is submitted for publication.

Weather and Air Surveillance Data

Official weather and air sampling data are seemingly quite accurate and reliable as compared with the host health data. These data have been improved greatly in recent years by the gradual substitution of objective measuring instruments for older methods of sensory judgment (e.g., rating of visibility).

Weather and air surveillance data are typically obtained by continuous automatic recording machines ranging from simple physical analog devices (e.g., the recording thermometer, the particulate matter air sampler, etc.) to complicated automatic chemical analyzers involving chemical reagents, filter columns, gas chromatography and spectrometry, etc. Many of these instruments are subject to drift and other artifactual perturbations which require frequent, expert and conscientious monitoring and recalibration. As in any human monitor system in which most of the time nothing unusual is happening, the corrections are often made in response to a subjective sensing of drift or of "bad" values, rather than according to a fixed protocol that is independent of the appearance of the data. Such monitoring could lead to artificial damping of extreme deviations or trends.

Gaps in daily data appear now and then due to mechanical failure, instrument maintenance, personnel shortages, etc. In the overall view, little information is lost in such gaps, but they do present serious problems to the statistical analyst and the computist.

The output of automatic weather and air recording devices is typically a continuous line graph on paper. The conversion from this analog of the air constituent to digital data is made by reading and, in some cases, measuring the graphs and recording numbers for the hourly reports. Although this task is not difficult and can be done with sufficient accuracy and precision for our purposes, it is subject to errors of many kinds, both biased and random, especially when done routinely day after day. Mechanical digital-analog converters or other nonhuman methods of obtaining a digital readout may in future eliminate some of this error. If the expense were warranted, a system could be developed in which the mechanically recorded, digitized, and stored data would some time later be fed into a computer for analysis. With a direct, continuous computer connection, the data could also serve for statistical quality control monitoring, with the computer itself actuating the necessary readjustments or recalibration. In any case, statistical quality control systems for continuous sample monitoring of the quality of weather and air data would probably be superior to current judgmental methods of controlling calibration and instrument error.

Analytical Models

a. Descriptive statistics. The early stages of study of any complex system usually consist in collection of observations, their assembly into apparently meaningful aggregates, and their condensation into numerical expressions or constants, such as means, standard deviations, regression equations, etc. Such expressions imply functional relationships among the variables they represent. If we have been fortunate in selecting an appropriate form of expression, they may have properties much the same as the observations. Once the form of these expressions has been chosen, the mathematical expressions may be selected so that they "fit" the observations better than any other expression of the same form but having different parameters. They are not, however, unique descriptions, since the choice of form of the expression, e.g., a linear relationship, was arbitrary from among many other forms that could also "fit" the observations. Such expressions can therefore be called descriptive. They say merely that the observations behave as if they had the stated functional relationship.

It cannot be properly inferred from such analyses, regardless of their statistical sophistication, that the functional relationship is that implied by the expression. Although a causal inference cannot be drawn, the writers recognize that many situations are encountered where investigators have to make decisions or draw conclusions and even to act on such incomplete information.

b. Causal inference. Having thus "guessed" a possible model or class of models for the way the variables observed in nature are related one to another, a series of controlled experiments may be conducted with the aim of excluding as unlikely whole classes of alternative models. As the experimental evidence accumulates we can narrow the range of uncertainty about which functional model is the correct one. Experiments can never "prove" that A causes B, but the range of possibilities can by experiment be narrowed to the point where we may infer that A causes B with a suitably small uncertainty.

In air pollution research we are still in the stage of "as if" description. Valid causal inference awaits experimental testing of cause and effect models based squarely on a plausible interpretation of what the mathematical expressions represent in the physical world. Sophisticated mathematical manipulations can lead to conclusions no better than the theory and hypotheses they rest upon.

Study methods available may be classified on a gradient from purely descriptive to purely inferential:

1. Record study. Analysis of extant records and observations to search for appearance of associations between environmental air factors and health indicators is limited to descriptive findings.

Experimental conditions of comparability of contrast groups, homogeneity of material, exclusion of irrelevant variables, etc., can only be simulated by statistical analysis, and they are limited to pseudo-controls on the variables that happen to have been recorded.

2. *Survey epidemiology*. Special surveys in search of more sensitive indicators of air pollution-health association can provide more specific co-variable information, data and groups not available in extant records, for purposes of strengthening the statistical controls on extraneous variables, thus bringing out relationships more clearly.

3. *Experimental epidemiology*. Limited segments of the total problem may be studied in the natural environment but with some real controls or treatment imposed in such a way as physically to eliminate some real controls or treatment imposed in such a way as physically to eliminate some of the extraneous variables.

> Examples of this stage of research are:
>
> direct household monitoring of air and disease in limited areas selected for high contrast in air pollution,
>
> direct study of the health of persons living in a high pollution environment compared with persons living in an adjacent filtered air environment, such as two wards of a hospital or nursing home, one with filtered air.

Harder to execute and thus less often contemplated are experiments in which subjects are randomly assigned to different treatments such as smoking or not smoking, wearing and not wearing a respiratory filter, moving or not moving to a contrasting community, etc.

4. *Laboratory experimentation*. For still more limited segments of the problem wholly artificial environments may be created in which to study health effects under controlled conditions. This may be done either for exploration or for investigation by testing quite specific hypotheses about possible underlying mechanisms. In either case the causal inferences are considerably stronger than in the previous case.

Simulation Models

Starting with known or inferred physical and chemical properties and relationships, a very simplified synthetic model of a cause and effect system may be built from elemental components. This in turn may be tested for agreement with experimental observations and then refined so as to simulate the real world more closely. Much can be learned about the real cause system by manipulating the various parameters of the model to explore, synthetically, the effects of the variables involved.

Simulation models may be physical analogs (e.g., a mechanical artificial lung), chemical analogs, electronic analogs, or purely mathematical analogs. (See references.)

References

Models and Analog Simulation

Bellman, R., Kalaba, R., and Jacquez, J.A., "Some Mathematical Aspects of Chemotherapy One-Organ Models." Rand Corp. Paper P-1550-RC (Nov. 1958).

Stacey, R.W., and Coulter, N.A., Jr., "Simulation of Human Physiological Systems." ASME Paper No. 60-AV-38, Dallas,(June 1960.)

Gold, A.J., "Possibilities for Simulation of Dynamic Physiology." ASME Paper No. 60-AV-40, Dallas, (June 1960).

Hoffman, J.G., Metropolis, N., and Gardner, V., "Digital Computer Studies of Cell Multiplication by Monte Carlo Methods." J. Nat. Cancer Inst., (Aug. 1956), pp.175-188.

Stacey, R.W. and Giles, F.M., "Computer Analysis of Arterial Properties." Circulation Research, 7:1031-8 (1959).

Grodins, F.S., "Intergrative Cardiovascular Physiology: A Mathematical Synthesis of Cardiac and Blood Vessel Hemodynamics." Quart. Rev. of Biology, 34:93-116 (June 1959).

Clynes, M., "Respiratory Control of Heart Rate: Laws Derived from Analog Computer Simulation." Trans. on Medical Electronics, ME-7, 2-14, (Jan. 1960).

Grodins, F.S., Avey, J.S., Schroeder, K.R., Norins, A.L., and Jones, R.W., "Respiratory Responses to CO_2 Inhalation: A Theoretical Study of a Nonlinear Biological Regulator." J. Applied Physiol. 7:238-308, (1954).

Dantzig, G., DeHaven, J., Cooper, I., Johnson, S., Deland, E., Kanter, H. and Samps, "A Mathematical Model of the Human External Respiratory System." Perspectives in Biol. and Med. 4:324 (1961).

In this section we will describe a number of mathematical and statistical techniques and review their application to air pollution research. Due to computation difficulties some of these methods have not been as widely used as desirable in the past. It should be noted that the availability of high speed computers removes a good deal of this difficulty. Most of the techniques described are used almost casually at U.C.L.A. and at other centers where there is easy access to high speed electronic computers.

To formalize the notation, suppose that we can measure several, say k, "response" variables and denote these by the letter Y with appropriate subscripts. Similarly we can measure several, say m, "dose" variables and denote these by the letter X with appropriate subscripts. All of these measurements can be made on several, say n, occasions, e.g., on several people, at several times, etc.

We will write

$$Y_{11}, Y_{12}, --, Y_{1k}; X_{11}, X_{12}, -- X_{1m}$$

for the first occasion observations, and generally

$$Y_{i1}, Y_{i2}, . ., Y_{ik}; X_{i1}, X_{i2}, . . ., X_{im}$$

for the observations on the ith occasion.

Curve and Surface Fitting

It is sometimes conjectured, believed, or hoped that some particular response variables can be related to the dose variables by a formula of specified type. For example we could ask whether a relation of the form

$$Y_i = a + bX_1 + cX_2$$

fitted the data, i.e., do the observed values of Y_{i1}, X_{i1}, and X_{i2} make the equations approximately true for some values of the coefficients a, b and c?

One approach is to determine values of a, b and c which make the fit as good as possible in some sense and then inspect the degree of fit attained. The sense of a good fit is usually least squares, that is the numbers $\hat{a}$, $\hat{b}$ and $\hat{c}$ which minimize the sum of the squared deviations,

$\varepsilon^2 = \Sigma(Y_{i1} - \hat{a} - \hat{b}X_{i1} - \hat{c}X_{i2})^2$ are used for a, b and c.

The degree of fit is sometimes measured by listing the deviations of the successive observations from the fitted equation, i.e.,

$$Y_{i1} - \hat{a} - \hat{b}X_{i1} - \hat{c}X_{i2},$$

and sometimes by computing one or more indices which presumably measure the fit. The most common such index is the correlation coefficient defined by

$$R^2 = 1 - \Sigma(Y_{i1} - \hat{a} - \hat{b}X_{i1} - \hat{c}X_{i2})^2 / \Sigma(Y_{i1} - \bar{Y}_1)^2$$

It should be noted that while a value of R^2 close to unity implies a good fit, a small value indicates only that the tried formula did not work and no information is given about the possible fit of other formulas. For example, one could devise a figure with an $R^2 = 0$, using a linear formula, but in which there obviously exists a relationship between the response and the dose.

A distinction should be made between values of R^2 which are statistically significant and those which are large enough to suggest an interesting degree of relationship. Sometimes the number $100(1-R^2)$ is quoted as the percent of variability in the response which remains after the formula is fitted--the amount of variability "unexplained" by the relationship. More appropriately, $\sqrt{100(1-R^2)}$ might be quoted since this gives the proportion of remaining variability in the original units of measurement.

Several formulas are somewhat routinely tried in attempts to describe or find relationships among variables. Some are:

(i) Linear relationships

$Y = A + BX$

$Y = A + BX_1 + CX_2$

or generally

$Y = A + B_1X_1 + B_2X_2 + . . + B_nX_n$

(ii) Functions linear in the coefficients but not linear in the variables

$Y = A + BX + CX^2$

$Y = A + BX_1 + CX_2 + DX_1^2 + EX_1X_2 + FX_2^2$

(or any polynomial function)

$Y = A + B \text{ Log } X$

$Y = A + B \cdot e^x$

$Y = A + B \sin X + C \cos X$

(or more generally, any finite series of this form)

(iii) Modified functions obtained from types above:

$\text{Log } Y = a + bX$, from $Y = A \cdot B^X$

$\text{Log } Y = a + B \text{ Log } X$, from $Y = A \cdot X^B$

$\frac{1}{Y} = A + BX$ from $Y = \frac{1}{A + BX}$

It should be noted that least square estimates using the modified formulas are not exactly equivalent to those using the original formulas. For most practical purposes they are certainly as good.

(iv) Special cases of more complicated formulas:

$Y = A + B \cdot C^x$

or, $\frac{1}{Y} = A + BC^x$ from $Y = \frac{1}{A + BC^x}$

$\text{Log } Y = a + bC^x$, $Y = A \cdot B^{cx}$

$Y = A + B_1 C_1^{x1} + B_2 C_2^{x2} + \ . \ . \ . + B_n C_n$

Cautions

1. From a scientific method point of view, the most desirable situation obtains when theory suggests some particular formula as appropriate. If it does not fit well, or if there is no such a priori formula, several formulas are routinely tried in the hope that a good fit will be found, and if so, that it will suggest some reasonable explanation of the way the variables behave.

2. It is a somewhat common, but to be discouraged, practice to attempt to "adjust" for a variable by using ratios of the other variables to the first. Neyman points out clearly the possible misinterpretations, and we feel categorically that other methods should be employed.

3. If the number of observations is not large, there may be correspondingly large sampling error and the hope of detecting an underlying relationship in the presence of other extraneous variables is small.

4. It must be explicitly recognized that a good relationship (e.g., R^2 close to 1) observed between variables in a survey does not necessarily imply a cause and effect relationship.

We are computing tables which will more clearly exhibit this effect and assist in planning the sample size required in surveys. One objective of a survey might be to estimate the value of R^2 as it exists in the population using the data from the sample. Although methods for measuring the statistical accuracy of this estimate for more than two variables are known in theory, the lack of adequate tables makes it generally ignored. Tables and graphs in preparation will give this accuracy with corresponding confidence intervals for cases of more than two variables. These will be of specific use in determining required samples sizes for survey studies.

In surveys as contrasted with experiments the investigation is at the mercy of the particular values of the dose variables that the observed situation chances to produce. Alternatively in experiments the investigator has the opportunity of selecting values of dose. This may increase the precision of the estimate of the population R^2. These remarks also apply when the hypothesis $R^2 = 0$ is to be tested. The tables referred to above will assist in determining the chance in a survey of detecting the falsity of the hypothesis in terms of possible alternative values of R^2.

Attention has been paid to the use of a priori reasonable causal relationships (paths) to reduce the number of parameters in a regression system.

These paths may be formulated as hypotheses to be tested or as involving interpretable parameters to be estimated. This technique probably has not been exploited as much as it might have been in air pollution research, especially where pollutant and meteorological variables are known to interact chemically or otherwise. In fact, general system analysis has been suggested for the atmosphere and possibly, with more access to large scale computers, such studies will become feasible.

Interrelationships among dose or among response variables. Fitting a formula as in the previous section does not allow for the possibility of almost as good fits with fewer variables. In fact, there is the possibility that simpler relationship may be concealed by unneeded extra variables. Alternatively, we might be searching for relationships among several variables rather than simply for one variable related to the rest. Component analysis is a method for attempting to detect linear relationships among sets of variables irrespective of whether they are to be treated as dependent or independent variables. This includes systematically choosing new coordinates or variables (components) which are linear functions of the original variables, which have the origin at the point representing the means of the original variables, and which in a sense indicate the best interrelationship among the variables, the second best, etc. The analysis includes the formulas giving the new variables (components) in decreasing order of importance in terms of the old and the proportion of the variability among the original variables which remains after each succeeding component is added. If this proportion gets close to zero, it follows that there are new variables remaining to be added.

Warning

Such relationships may not show up specifically among the new variables.

The problem of interpreting these components may be difficult. First the source of data should be reinspected to see if any artificial relationships have been introduced, for example, if the difference of two variables, such as high temperature minus low temperature may have been used as a third variable, or if the mean temperature over several days may have been added as a variable along with the daily temperatures. In some cases, well known or seemingly related variables might have been included such as temperature at two closely situated weather stations.

Secondly, after removing variables artificially introduced, if any are found, an attempt can be made to systematically study subsets of variables in the hope that relationships exist among fewer variables which in turn may be easier to identify. This is particularly convenient to do if one has access to a computer.

Thirdly, it may be possible to explain rationally the formulas giving the principal components. Note that since only linear components are used and since the new origin is at the means of the original variables, the method will not be powerful in the presence of relationships which are not linear. In this respect, it has the same failing as linear regression analysis. Needed, especially when large amounts of data are available, is a method of approach which indicates whether any relationships exist rather than whether some specific one does or does not.

Extensions of component analysis would include factor analysis and canonical analysis methods.

Appendix I

LOS ANGELES HOSPITAL ADMISSIONS STUDY, 1958

Standard Nomenclature Code	Codes Used in This Study	Diagnosis
		Major Respiratory
340-100	11	Acute tracheitis due to infection
350-100	12	Bronchitis, acute
350-100.0	13	Bronchitis, chronic
350-100.6	14	Bronchiectasis due to infection
350-yx6	15	Bronchiectasis due to undetermined cause
350-3	16	Bronchitis due to infection
350-390	17	Asthma
353-100	18	Bronchiolitis, acute
360-100	19	Pneumonia, lobar
361-190	20	Bronchopneumonia
361-y10	21	Bronchopneumonia due to undetermined cause
362-9x6	22	Emphysema, pulmonary, due to undetermined cause
		Miscellaneous Major Respiratory
340-100.4	23	Stenosis of trachea due to infection
350-100.4		Stenosis of bronchus due to infection
360-100.8		Edema, pulmonary, due to infection
362-100.4		Atelectasis due to infection
362-797		Senile emphysema
3031-100		Laryngotracheitis, acute
		Minor Respiratory
300-100	24	Common cold
310-100		Rhinitis, acute (diagnosed ordinarily as common cold)
310-392		Allergic rhinitis (hay fever)
316-940		Middle turbinate-hypertrophic rhinitis due to unknown cause
318-100		Nasopharyngitis
320-100		Sinusitis, pansinusitis, acute
320-190		Sinusitis, nonpurulent, chronic
321-100		Sinusitis, maxillary, acute
322-100		Sinusitis, frontal, acute
330-100		Laryngitis, acute
330-100.0		Laryngitis, chronic, due to infection
330-994		Mucous polyp of larynx, simple
331-100		Epiglottitis
3001-100		Acute diffuse upper respiratory infection

Appendix I

LOS ANGELES HOSPITAL ADMISSIONS STUDY, 1958

Standard Nomenclature Code	Codes Used in This Study	Diagnosis
		Major Respiratory
340-100	11	Acute tracheitis due to infection
350-100	12	Bronchitis, acute
350-100.0	13	Bronchitis, chronic
350-100.6	14	Bronchiectasis due to infection
350-yx6	15	Bronchiectasis due to undetermined cause
350-3	16	Bronchitis due to infection
350-390	17	Asthma
353-100	18	Bronchiolitis, acute
360-100	19	Pneumonia, lobar
361-190	20	Bronchopneumonia
361-y10	21	Bronchopneumonia due to undetermined cause
362-9x6	22	Emphysema, pulmonary, due to undetermined cause
		Miscellaneous Major Respiratory
340-100.4	23	Stenosis of trachea due to infection
350-100.4		Stenosis of bronchus due to infection
360-100.8		Edema, pulmonary, due to infection
362-100.4		Atelectasis due to infection
362-797		Senile emphysema
3031-100		Laryngotracheitis, acute
		Minor Respiratory
300-100	24	Common cold
310-100		Rhinitis, acute (diagnosed ordinarily as common cold)
310-392		Allergic rhinitis (hay fever)
316-940		Middle turbinate-hypertrophic rhinitis due to unknown cause
318-100		Nasopharyngitis
320-100		Sinusitis, pansinusitis, acute
320-190		Sinusitis, nonpurulent, chronic
321-100		Sinusitis, maxillary, acute
322-100		Sinusitis, frontal, acute
330-100		Laryngitis, acute
330-100.0		Laryngitis, chronic, due to infection
330-994		Mucous polyp of larynx, simple
331-100		Epiglottitis
3001-100		Acute diffuse upper respiratory infection

Standard Nomenclature Code	Codes Used in This Study	Diagnosis
		Major Cardiac
400-533	01	Hypertensive cardiovascular disease
400-535	02	Cardiovascular disease due to hypertension of lesser circulation
410-535.6		Enlargement of heart due to pulmonary disease
410-516	03	Arteriosclerotic heart disease
410-5x7.5	04	Fibrosis of myocardium due to infarction
430-516.7		Infarction of myocardium due to arteriosclerotic coronary thrombosis
430-512.7		Infarction of myocardium due to embolism
430-515.7		Infarction of myocardium from coronary thrombosis due to arteritis
430-5x7.6		Fibrosis of myocardium due to infarction
430-9x7		Infarction of myocardium due to unknown cause
410-522	05	Passive congestion of heart
430-522.6		Fibrosis of myocardium due to passive congestion
430-955		Fibrosis of myocardium due to unknown cause
410-9x6	06	Enlargement of heart due to unknown cause
		Miscellaneous Major Cardiac
404-900.4	07	Mediastinopericarditis due to unknown cause
410-y00		Heart disease, undiagnosed
410-4x4		Cardiac tamponade
410-100		Acute bacterial carditis
410-402.x		Cardiac arrest
410-43x.6		Hypertrophy of heart due to overstrain
420-100		Acute bacterial pericarditis
420-100.3		Pneumopericardium due to infection
430-100		Acute bacterial myocarditis
430-102		Myocarditis due to scarlet fever
430-160		Myocarditis due to virus
430-168		Myocarditis due to influenza
430-930		Acute isolated myocarditis due to unknown cause
430-950		Myocarditis of unknown cause

Standard Nomenclature Code	Codes Used in This Study	Diagnosis
		Less Prevalent Cardiac
408-584.x	08	Carotid sinus syncope (syndrome)
431-x21		Atrial premature contraction due to unknown cause
431-x22		Atrial paroxysmal tachycardia due to unknown cause
431-x23		Paroxysmal atrial flutter due to unknown cause
431-x25		Paroxysmal atrial fibrillation due to unknown cause
434-x41		Ventricular premature contractions due to unknown cause
434-x42		Ventricular paroxysmal tachycardia due to unknown cause
434-x44		Ventricular standstill due to unknown cause
434-x45		Ventricular fibrillation due to unknown cause
440-580		Neurogenic arrhythmia
440-584		Vagal bradycardia
440-585		Vagal arrhythmia
440-590		Reflex arrhythmia
441-x11		Sinus arrest due to unknown cause
441-x12		Sinus arrhythmia due to unknown cause
443-x32		Paroxysmal tachycardia originating in the atrioventricular node due to unknown cause

A-1

Appendix II

Epidemiological Studies of Obstructive Ventilatory Diseases III

Maximal Expiratory Flow Tests in a Population Sample with An Application of Component Analysis*

by

Michael Zwerdling

John R. Goldsmith, M.D.

and

Frank Massey Jr., Ph.D.

DRAFT for Possible Publication*

---

*Within these drafts, the enumeration of tables and figures is distinct from that in the remainder of the report.

Introduction

Several tests of ventilatory obstruction are used in the study of obstructive ventilatory disease. It would be valuable to determine the relationships of these tests among themselves and to other criteria of obstructive disease, such as dyspnea, cough, sputum, physical examination, and roentgenologic examination. Since test values are known to be influenced by age, sex, and physique, these variables should also be included in any study of the relationships among test values. Finally, it is thought that cigarette smoking has some influence on the presence or extent of obstructive disease; therefore, smoking should be a variable to be included in any analysis. There are, thus, four classes of variables in this study: Variables reflecting (1) Physiological Test Results, (2) Age, Sex and Physique, (3) Clinical Findings, and (4) Exposures to Cigarette Smoking. This paper reports methods and results of examining the relationships among these measures and classes of variables in a sample of a general population.

Investigators in the medical sciences frequently work with a number of interrelated variables when they investigate the intricate nature of the human body. Recently, new statistical methods for investigating these complex relationships have been developed along with electronic computer facilities for applying these methods. It was felt desirable to apply these methods to a set of data containing some known interrelationships between variables. This study concerns the application of component analysis to a set of data collected at the Framingham Heart Disease Epidemiology Program of the National Heart Institute. The analysis was undertaken partly to become familiar with this technique and to ascertain the effect of certain alterations of the data input on the results of analysis, and partly to portray the relationships among the variables studied. The analysis also helps to determine how many properties of the respiratory system are measured by this battery of function tests and how these properties are related to the anthropometric, health and smoking variables which were also determined. Criteria for determining the statistical significance of the analytical results are not readily available; therefore, conclusions about the relationship of variables and the number of "components" of pulmonary function available in the data must be considered tentative.

Methods for Obtaining Data

The Framingham Heart Disease Epidemiology Program involves a biennial medical history and physical examination of a population sample of the residents of Framingham, Massachusetts. During the fifth biennial examination, a battery of pulmonary function tests was performed on a subsample of this group. The subsample consisted of all persons, examined during a period of weeks, who had complete data.

The tests and measurements were performed in the following order: height, weight, and strength of hand grip were obtained first, followed by a maximal expiratory flow test using an indicating device (Puffmeter). The maximal flow test was performed with the subject standing. Nasal occlusion was not used in any test. The subject was told to take in as much air as possible and to blow it through the instrument. The device consists of a flow-to-pressure transducer and a calibrated indicating pressure gage. Ten tests were obtained from each subject within a few minutes. The mean of the highest sequence of three tests was used to represent the subject's performance.

Vital capacity was tested using a standard Collins spirometer. The highest of two values was used. No attempt was made to control the rate of expiration, in contrast to the data from the Krogh spirometer, discussed below.

Three replications of a forced expiratory flow test using a Krogh spirometer with a Microsyn transducer were performed by another technician. The subjects were instructed to expel the air as fast as possible while standing. The output from the transducer was recorded by one channel of a Sanborn Twin Viso Cardiette at a paper speed of 50 millimeters per second. The output (volume) signal of the first channel was differentiated, and the differentiated signal (flow) was recorded on the second channel. Of the three tests, the one with the greatest flow was recorded and analyzed. The volume expelled at intervals of 0.2, 0.5, 0.75, and 1.0 second, as well as the total vital capacity, were determined on this tracing. These volumes expressed as a percent of the total vital capacity were calculated also.

The mid-expiratory time is determined from this same tracing by measuring the time which is required to expel the middle half of the expiratory vital capacity. When ahlf the vital capacity is divided by this time, the result is the maximal mid-expiratory flow.

A history was taken and physical examination performed by an examining physician on the staff of the program. A fourteen by seventeen inch chest roentgenogram was taken under uniform conditions and read by a roentgenologist.

For 43 males and 63 females between the ages of 37 and 67, the following information, which can be classified into the four categories of variables previously mentioned, was obtained by the trained nursing, medical, and technical staff members:

(These numbers refer to the
entries used in the tables.
Units used may be found in
Table I.)

A. Physiological Test Results

Mid-Expiratory Time	8
Puffmeter	9
Maximum Expiratory Flow (Differentiated Circuit)	10
0.2 Second Forced Expiratory Volume (FEV$_{0.2}$)	11
0.5 Second Forced Expiratory Volume (FEV$_{0.5}$)	12
0.75 Second Forced Expiratory Volume (FEV$_{0.75}$)	13
1.0 Second Forced Expiratory Volume (FEV$_{1.0}$)	14
Forced Vital Capacity (FVC)	15
Spirometric Vital Capacity (VC)	16
0.2 Second Forced Expiratory Volume/Forced Vital Capacity (FEV$_{0.2\%}$)	17
0.5 Second Forced Expiratory Volume/Forced Vital Capacity (FEV$_{0.5\%}$)	18

0.75 Second Forced Expiratory Volume/Forced Vital Capacity ($FEV_{0.75\%}$)	19
1.0 Second Forced Expiratory Volume/Forced Vital Capacity ($FEV_{1.0\%}$)	20
Maximum Mid-Expiratory Flow	21

B. Variables Reflecting Age, Sex, and Physique

Age	1
Height	2
Weight	3
Body Surface Area (BSA) from Nomograms based on the formula: $BSA = W^{0.425} \times H^{0.725} \times C$ (W equals weight, H equals height, and C is a constant which varies with the units of measurement)	4
Chest Diameter	5
Strength of Hand Grip	6
Obesity - weight of individual divided by the study population mean weight for height and sex multiplied by 100	7

C. Clinical Findings	23

The following code categories represented the variables representing clinical findings:

Category 0

Absence of any of the findings listed below

Category 1

Unanticipated or unclassified conditions which may influence respiratory function. Included were pregnancy, chest deformity, etc. Subjects with such conditions were not included in the analysis.

Category 2

a. Previous history of hay fever, asthma, tuberculosis, chest surgery, or, within the previous two years, bronchitis or pneumonia;

b. No current symptoms of cough, sputum, shortness of breath;

c. No abnormal findings on physical examination of the chest;

d. No roentgenographic evidence of diffuse parenchymal or pleural disease.

Category 3

Current symptoms of cough, sputum or shortness of breath but no abnormal physical or roentgenologic findings.

Category 4

Physical examination revealed wheezing, serious chest deformity, or rales but no current symptoms of cough, sputum, or shortness of breath.

Category 5

Roentgenogram interpreted as showing fibrosis, pleural deformity, emphysema or diffuse nodularity but no current symptoms of cough, sputum, or shortness of breath.

Category 6

Current symptoms of cough, sputum, or shortness of breath were present together with abnormal physical or roentgenographic findings meeting the specifications for Categories 4 or 5.

This category code was intended to provide an approximation in numerical sequence to the likelihood of diffuse pulmonary disease being present. A reasonably good approximation probably exists for categories 0, 2, 3, and 6.

D. Exposures to Cigarette Smoking 22

Code

0 = not a cigarette smoker
1 = 20 or fewer cigarettes per day
2 = 21-30 cigarettes per day
3 = 31 or more cigarettes per day

From the nature of the measurements made, it is obvious that some of the variables are interrelated. For example, body surface area (No. 4) and obesity (No. 7) are dependent on weight (No. 3). In addition, body surface is dependent on height (No. 2). Variables No. 9 through No. 14 and No. 21 are all estimates of respiratory flow and only two of these, No. 9 and any one other, were obtained by totally independent methods. There are two independent estimates of vital capacity (No. 15 and No. 16); however, No. 15 represents a maximally rapid test. Five of the variables, Nos. 17 through 21, are quotients of two of the others.

Statistical Methods

Component Analysis is a statistical method for investigating the relationships among a set of variables. The objective of the analysis is to determine whether the relationship among the observed variables can be accounted for adequately by a smaller number of new variables, corresponding to the "principal components". The principal components are not necessarily identifiable in the sense of corresponding to any observable physical measurement. However, by analyzing the relationship of the observed variables to the principal components, an approximation to the meaning of a given component may be obtained. In some cases, it may then be possible to devise a method which will better estimate the component.

In order to explain the nature of component analysis in greater detail, explanations in both geometric and symbolic terms will be given.

If two variables are measured on a number of individuals, the results may be plotted as the points in the plane of a scatter diagram. If a third variable is added, the results might be plotted as points in a three dimensional space. By analogy, if 23 variables are measured on 43 men, the results might be represented as 43 points in a 23 dimensional space. Although it is difficult to think in terms of 23 dimensions, one might picture the observations as a swarm of points which from any vantage point in 23 dimensional space may be represented by an ellipsoid body, or egg; the method of principal components in effect determines the possible axes (up to 23) of this hypothetical ellipsoid body or egg. The first component corresponds to the longest axis of the egg-shaped swarm of points and identifies the location of each of the 43 points with respect to this axis. The second component corresponds to the axis drawn at right angles to the first component in the longest remaining axis. In all, the number of components may be equal to the number of dimensions; that is, to the number of original variables. In practice, however, due to the calculations used, since each component successively removes the maximum amount of remaining variability, a useful description of the observed variation may be obtained by locating the first few principal components. If the approximation is an adequate representation, then the original 23 variables have been effectively reduced to a much smaller number.

In symbolic notation p variables which may be interrelated are observed on n individuals. The problem is to express the data in terms of m, a number less than p, uncorrelated variables. The first new variable (component) is chosen such that it represents as large a proportion of the combined variance of all p variables as possible; the second component is uncorrelated with the first and has as large a proportion of the remaining variance as possible, and so on. Thus, the data are transformed to a form in which each component is uncorrelated with any other and each component accounts for as much of the remaining variation as possible, in descending order. If two or three of the components account for, say, 85 percent of the variation, then a reasonable approximation to a representation of the data has been achieved.

With the availability of an electronic computer and the appropriate programs, we could easily compute many repetitions of a given problem; therefore, several variations on this general problem were studied; namely, a) analysis with males and females separately and with both groups combined, b) analysis using twelve measures derived from a single forced expiratory flow tracing, and, later, two measures derived from the same data, and c) a small sample followed by an augmented sample. These three variations represent comparisons of analysis upon groups with 1) different degrees of homogeneity, 2) different degrees of replication of interrelated variables, and 3) different sample sizes.

Results

Appendix A shows the product-moment correlation coefficients among all the variables separately by sex. The statistical criteria for significance can be greatly influenced by the use of variables which are quotients or which are arbitrary numerical codes rather than measurements. The possible significance of the correlations and the significance of the differences between males and females for correlations of such variables (quotients or codes) cannot be readily determined. Table I shows the means and standard deviations for the 23 variables for 43 males and 63 females.

Component analyses were performed on the data for the 43 males and 63 females, as well as for the combined group totaling 106. The analyses were repeated for the same samples, using two subsets of nine variables in each analysis. Finally, an analysis was performed using the two subsets of nine variables with a sample including the 106 persons in the original analysis, plus an additional group for whom complete information was available for one set of the nine variables and a still larger group with information on the other set of nine variables, but not for the 23 variables.

From the analyses using 23 variables, it was possible to determine the relationships of the components to the original variables and to compare the relationships for males and females. Combining males and females gave an indication of the effects on the component analysis of combining two heterogeneous populations. The results of these procedures are shown in Table II and in Section A of Figures 1, 2, and 3.

In all three analyses, the predominant pattern is that the first and second component are associated with pulmonary function tests and the third is associated with body weight. Approximately 73 percent of the total variance in each sample is accounted for by these three components. The significant points appear to be: 1) the 14 pulmonary function tests appear to be measuring only two independent aspects of lung function, 2) height, hand grip strength, and age are the age-physique variables most closely associated with the two components reflecting lung function, and 3) the relationships of variables to components is fairly similar in males, females, and both groups combined, although there are some dissimilarities. It should be pointed out that there is no statistical technique known to us for assessing the significance of the differences observed in component analyses of two groups.

Tables III and IV and Sections B and C of Figures 1, 2, and 3 show the relationships between variables and components for the same persons discussed above using two different subsets of nine variables, differing only in the choice of FEV_t. The proportion of the variance accounted for by the first three components is approximately the same as with the 23 variables. For males and females considered separately, the pattern of two components associated with pulmonary function and a third component associated with body weight is repeated. The major differences appear to be: 1) the second component is more strongly reflected in the percent vital capacity, and 2) the anthropometric measurements are more strongly associated with the first two components. When males and females are combined, using the nine variables, the third component no longer appears to be associated with body weight, apparently being more closely related to age and percent capacity.

One would expect no change in the relationship of variables to components when sample size is increased with no change in the variables considered, provided both samples were randomly drawn from the same universe. The reason for this expectation is that one would expect only random changes in the observed correlations between variables, and it is from these correlations that the results of component analysis are derived. The results of the component analyses are shown in Tables V and VI and Sections D and E of Figures 1, 2 and 3. For both sexes combined, the results are comparable to the corresponding analysis for the smaller sample. For males, the first two components are similar to those in the smaller sample, but the third component is more closely associated with clinical findings than with body weight. For females, the first and third component are associated with pulmonary function, while the second component is associated with body weight in the larger augmented sample. The reason for this apparent change is that more of the total variance is associated with the weight component than is associated with the pulmonary function component for this sample. The difference in the amount of variance explained by the two components is very small, as can be seen in Table VI.

It is important that the general nature of the relationships of two components associated with pulmonary function and a third component associated with body weight seems fairly constant in each of the sets of data analyzed.

The possibility of extracting additional components beyond the third was explored; but since the proportion of variance accounted for becomes progressively smaller and the degree of association with pulmonary function tests is low, little would be gained.

Discussion:

In Calculating $FEV_{t\%}$, a ratio of two of the variables, FEV_t and VC, is used. FEV_t and VC tend to be fairly closely correlated, and therefore the ratio of one to the other tends to be a constant, uncorrelated with either of the original variables. Since the first principal component in most of the analyses is closely related to both FEV_t and VC, it is unlikely that $FEV_{t\%}$ would be closely related to the first component. The point to note is that the second component, which is related to $FEV_{t\%}$, may not be measuring an independent attribute of lung function, but may merely be a result of the numerical procedure and the considerations described above.

SUMMARY

1. Data reflecting: a) physiological test results, b) age, sex and physique, c) clinical findings and d) exposures to cigarette smoking were collected on a subset of the population sample being studied.

2. To determine the feasability of applying computer technology and techniques of multivariate analysis, the data were subjected to component analysis, a method by which one attempts to transform a large number of interrelated variables into a smaller number of uncorrelated variables.

3. Analyses were run on several subsets of variables and for several samples. In practically all cases the first three components were associated with two aspects of pulmonary function and with body weight and accounted for approximately three fourths of the total variance.

ChD
071061

A-12

Table 1

MEAN AND STANDARD DEVIATION FOR 23 VARIABLES MEASURED ON 43 MALES AND 63 FEMALES

NUMBER	VARIABLE	MALE		FEMALE	
		Mean	Standard Deviation	Mean	Standard Deviation
1	Age (years)	52.7	7.6	50.8	8.2
2	Height (inches)	67.1	2.5	63.0	2.7
3	Weight (pounds)	166.5	19.9	142.6	23.3
4	Body surface area (square meters)	1.87	0.12	1.67	0.14
5	Chest diameter (inches)	19.6	1.8	16.1	1.8
6	Dynamometer (strength of hand grip)	454.4	78.0	259.4	54.8
7	Obesity[1]	102.4	12.5	102.8	16.2
8	Mid-expiratory time (seconds)	.68	.29	.60	.20
9	Puffmeter (liters per minute)	351.2	65.5	292.5	47.5
10	Maximum expiratory flow (differentiated circuit – liters per minute)	644.4	175.2	522.4	170.6
11	0.2 second forced expiratory volume (liters)	1.18	0.36	0.87	0.26
12	0.5 second forced expiratory volume (liters)	2.20	0.53	1.68	0.44
13	0.75 second forced expiratory volume (liters)	2.61	0.58	2.01	0.52
14	1.0 second forced expiratory volume (liters)	2.85	0.60	2.18	0.54
15	Forced vital capacity (liters)	3.50	0.69	2.65	0.60
16	Spirometric vital capacity (liters)	3.61	0.62	2.74	0.60
17	0.2 second vol./forced vital capacity	0.34	0.09	0.33	0.09
18	0.5 second vol./forced vital capacity	0.63	0.11	0.64	0.10
19	0.75 second vol./forced vital capacity	0.75	0.11	0.76	0.09
20	1.0 second vol./forced vital capacity	0.82	0.10	0.82	0.08
21	Maximum mid-expiratory flow (liters per minute)	180.0	72.0	146.3	51.8
22	Smoking[2]	1.07	0.85	0.44	0.66
23	Respiratory disease category[3]	3.26	2.14	1.87	1.95

[1] (weight of individual/mean weight for height and sex) x 100.

[2] 0 = nonsmoker; 1 = 20 or fewer cigarettes daily; 2 = 21-30 cigarettes daily; 3 = 31 or more cigarettes daily.

[3] 0 = normal; 2 = previous history only; 3 = present symptoms only; 4 = positive physical examination without symptoms; 5 = positive X-ray without symptoms; 6 = positive X-ray or physical examination with present symptoms.

A-13

Table 2

PERCENT OF VARIANCE ASSOCIATED WITH EACH OF THE FIRST THREE PRINCIPAL COMPONENTS FOR 23 VARIABLES, BY SEX

NUMBER	VARIABLE	BOTH SEXES N = 106					MALE N = 43					FEMALE N = 63				
		Components					Components					Components				
		Total	I	II	III	Residual	Total	I	II	III	Residual	Total	I	II	III	Residual
	Total variance	100.0	40.6	21.1	12.6	25.7	100.0	40.6	19.0	13.9	26.5	100.0	37.7	20.5	14.4	27.4
1	Age	100.0	7.8	1.4	14.4	76.4	100.0	26.0	8.4	2.6	63.0	100.0	13.0	28.1	3.6	55.3
2	Height	100.0	42.3	30.7	0.2	26.8	100.0	4.4	44.9	7.3	43.4	100.0	37.2	25.0	2.3	35.5
3	Weight	100.0	20.3	10.2	62.4	7.1	100.0	4.8	0.8	90.3	4.1	100.0	4.0	1.4	90.3	4.3
4	Body surface area	100.0	36.0	21.2	36.0	6.8	100.0	6.8	4.4	77.4	11.4	100.0	17.6	7.3	68.9	6.2
5	Chest diameter	100.0	1.7	1.0	67.2	30.1	100.0	6.8	16.0	40.0	37.2	100.0	0.2	7.8	57.8	34.2
6	Dynamometer	100.0	44.9	23.0	2.6	29.5	100.0	19.4	16.8	19.4	44.4	100.0	22.1	17.6	2.0	58.3
7	Obesity[1]	100.0	0.6	0.0	69.8	29.6	100.0	1.2	16.8	65.6	16.4	100.0	0.1	0.1	88.4	11.4
8	Mid-expiratory time	100.0	14.4	65.6	0.0	20.0	100.0	43.6	33.6	1.4	21.4	100.0	29.2	46.2	0.3	24.3
9	Puffmeter	100.0	67.2	0.2	0.6	32.0	100.0	65.6	0.2	1.7	32.5	100.0	54.8	4.4	0.3	40.5
10	Maximum expiratory flow	100.0	60.8	2.3	0.0	36.9	100.0	57.8	0.1	1.2	40.9	100.0	57.8	4.8	1.0	36.4
11	$FEV_{0.2}$	100.0	82.2	1.7	0.3	15.8	100.0	82.8	0.1	0.5	16.6	100.0	74.0	4.4	0.3	21.3
12	$FEV_{0.5}$	100.0	90.8	0.0	2.3	6.9	100.0	90.8	3.2	1.2	4.8	100.0	86.5	0.0	1.0	12.5
13	$FEV_{0.75}$	100.0	89.0	1.0	4.8	5.2	100.0	84.6	10.2	0.6	4.6	100.0	86.5	2.9	2.3	8.3
14	$FEV_{1.0}$	100.0	84.6	4.0	5.8	5.6	100.0	77.4	16.8	0.5	5.3	100.0	81.0	8.4	2.6	8.0
15	FVC	100.0	59.3	27.0	7.3	6.4	100.0	29.2	60.8	0.4	9.6	100.0	54.8	31.4	1.5	12.3
16	VC	100.0	62.4	24.0	4.0	9.6	100.0	34.8	49.0	0.0	16.2	100.0	54.2	34.4	0.5	10.9
17	$FEV_{0.2\%}$	100.0	16.0	53.3	5.3	25.4	100.0	47.6	32.5	0.1	19.8	100.0	9.6	67.2	0.5	22.7
18	$FEV_{0.5\%}$	100.0	20.7	68.1	2.9	8.3	100.0	51.1	41.0	0.8	7.1	100.0	22.1	68.0	0.1	9.8
19	$FEV_{0.75\%}$	100.0	24.0	67.7	0.6	7.7	100.0	51.8	39.7	0.5	8.0	100.0	34.4	54.8	0.5	10.3
20	$FEV_{1.0\%}$	100.0	23.5	60.8	0.2	15.5	100.0	47.6	37.2	0.2	15.0	100.0	34.8	43.6	1.0	20.6
21	Maximum mid-expiratory flow	100.0	70.6	11.6	0.8	17.0	100.0	75.3	3.6	0.2	20.9	100.0	77.4	5.3	0.8	16.5
22	Smoking[2]	100.0	9.6	5.3	0.0	85.1	100.0	2.0	0.1	1.7	96.2	100.0	0.5	3.2	5.3	91.0
23	Disease category[3]	100.0	2.0	3.6	2.9	91.5	100.0	21.2	0.1	5.8	72.9	100.0	15.2	5.8	0.0	79.0

[1] (weight of individual/mean weight for height and sex) x 100.

[2] 0 = nonsmoker; 1 = 20 or fewer cigarettes daily; 2 = 21-30 cigarettes daily; 3 = 31 or more cigarettes daily.

[3] 0 = normal; 2 = previous history only; 3 = present symptoms only; 4 = positive physical examination without symptoms; 5 = positive X-ray without symptoms; 6 = positive X-ray or physical examination with present symptoms.

Table 3

PERCENT OF VARIANCE ASSOCIATED WITH EACH OF THE FIRST THREE PRINCIPAL COMPONENTS FOR 9 VARIABLES, BY SEX

NUMBER	VARIABLE	BOTH SEXES N = 106					MALE N = 43					FEMALE N = 63				
		Components					Components					Components				
		Total	I	II	III	Residual	Total	I	II	III	Residual	Total	I	II	III	Residual
	Total variance	100.0	45.2	16.7	14.0	24.1	100.0	40.4	18.9	14.0	26.7	100.0	43.3	16.5	11.5	28.7
1	Age	100.0	12.3	14.4	46.8	26.5	100.0	44.9	0.3	1.7	53.1	100.0	41.0	16.8	9.6	32.6
2	Height	100.0	70.6	11.6	0.6	17.2	100.0	22.0	50.4	4.0	23.6	100.0	65.2	1.0	0.4	33.4
3	Weight	100.0	30.8	7.3	3.6	58.3	100.0	11.6	2.9	51.8	33.7	100.0	9.0	0.5	77.4	13.1
6	Dynamometer	100.0	69.7	9.6	0.1	20.6	100.0	49.0	13.0	12.3	25.7	100.0	50.4	5.3	3.6	40.7
9	Puffmeter	100.0	67.7	6.8	2.0	23.5	100.0	56.3	10.2	10.9	22.6	100.0	65.6	5.8	0.0	28.6
12	FEV$_{0.5}$	100.0	70.1	13.0	4.4	12.5	100.0	74.0	9.0	5.3	11.7	100.0	51.0	31.4	4.0	13.6
16	VC	100.0	82.8	0.8	4.0	12.4	100.0	62.4	14.8	13.7	9.1	100.0	81.0	1.7	1.4	15.9
18	FEV$_{0.5\%}$	100.0	2.0	42.3	4.6	51.1	100.0	18.1	62.4	0.0	19.5	100.0	0.0	86.5	0.0	13.5
23	Category[1]	100.0	0.4	44.9	17.6	37.1	100.0	27.0	8.4	26.0	38.6	100.0	25.5	0.0	4.0	70.5

---

[1] 0 = normal; 2 = previous history only; 3 = present symptoms only; 4 = positive physical examination without symptoms; 5 = positive X-ray without symptoms; 6 = positive X-ray or physical examination with present symptoms.

A-15

Table 4

PERCENT OF VARIANCE ASSOCIATED WITH EACH OF THE FIRST THREE PRINCIPAL COMPONENTS FOR 9 VARIABLES, BY SEX

NUMBER	VARIABLE	BOTH SEXES N = 106					MALE N = 43					FEMALE N = 63				
		Components					Components					Components				
		Total	I	II	III	Residual	Total	I	II	III	Residual	Total	I	II	III	Residual
	Total variance	100.0	46.6	17.6	11.0	24.8	100.0	45.6	13.6	11.1	29.7	100.0	41.7	17.3	14.1	26.9
1	Age	100.0	14.4	29.2	30.3	26.1	100.0	46.2	0.3	1.4	52.1	100.0	42.3	0.4	24.0	33.3
2	Height	100.0	74.0	9.6	0.8	15.6	100.0	27.0	47.6	0.3	25.1	100.0	64.9	5.8	0.5	28.8
3	Weight	100.0	29.2	10.9	2.0	57.9	100.0	11.2	0.1	59.7	29.0	100.0	6.8	29.2	51.8	12.2
6	Dynamometer	100.0	67.2	10.2	0.0	22.6	100.0	49.0	6.3	16.8	27.9	100.0	51.0	12.3	0.8	35.9
9	Puffmeter	100.0	64.8	4.4	6.8	24.0	100.0	54.8	10.9	14.4	19.9	100.0	62.4	1.7	2.9	33.0
14	$FEV_{1.0}$	100.0	78.3	4.4	0.3	17.0	100.0	79.2	0.4	7.3	13.1	100.0	72.3	10.2	0.0	17.5
16	VC	100.0	86.5	0.2	2.9	10.4	100.0	67.8	15.2	9.0	8.0	100.0	84.6	0.1	2.3	13.0
20	$FEV_{1.0\%}$	100.0	4.0	38.4	49.0	8.6	100.0	16.0	56.3	1.0	26.7	100.0	6.3	59.3	16.8	17.6
23	Category[1]	100.0	0.6	51.8	6.8	40.8	100.0	26.0	20.3	16.8	36.9	100.0	21.6	2.9	0.3	75.2

[1] 0 = normal; 2 = previous history only; 3 = present symptoms only; 4 = positive physical examination without symptoms; 5 = positive X-ray without symptoms; 6 = positive X-ray or physical examination with present symptoms.

Table 5

PERCENT OF VARIANCE ASSOCIATED WITH EACH OF THE FIRST THREE PRINCIPAL COMPONENTS FOR 9 VARIABLES, BY SEX AUGMENTED SAMPLE

NUMBER	VARIABLE	BOTH SEXES N = 170					MALE N = 68					FEMALE N = 102				
		Total	Components I	Components II	Components III	Residual	Total	Components I	Components II	Components III	Residual	Total	Components I	Components II	Components III	Residual
	Total variance	100.0	45.6	15.8	13.8	24.8	100.0	41.4	17.9	12.0	28.7	100.0	40.2	16.9	13.2	29.7
1	Age	100.0	9.5	18.0	38.1	34.4	100.0	37.2	1.1	6.2	55.5	100.0	38.4	15.1	8.1	38.4
2	Height	100.0	68.7	12.3	0.3	18.7	100.0	43.1	26.5	7.3	23.1	100.0	46.1	0.7	24.1	29.1
3	Weight	100.0	38.7	3.8	5.7	51.8	100.0	28.0	8.7	26.3	37.0	100.0	9.5	13.8	44.1	32.6
6	Dynamometer	100.0	70.1	9.2	0.0	20.7	100.0	52.4	15.1	0.6	31.9	100.0	46.9	9.7	6.8	36.6
9	Puffmeter	100.0	66.7	6.7	0.9	25.7	100.0	50.7	20.4	4.5	24.4	100.0	61.4	5.5	2.4	30.7
12	$FEV_{0.5}$	100.0	72.1	10.7	4.5	12.7	100.0	66.4	19.8	1.4	12.4	100.0	59.3	21.8	3.2	15.7
16	VC	100.0	83.9	0.3	3.9	11.9	100.0	67.3	0.9	21.4	10.4	100.0	81.9	3.4	0.4	14.3
18	$FEV_{0.5}\%$	100.0	0.7	41.8	50.4	7.1	100.0	9.5	68.2	6.1	16.2	100.0	0.2	81.3	2.7	15.8
23	Category[1]	100.0	0.4	39.0	20.3	40.3	100.0	18.4	0.3	33.9	47.4	100.0	17.8	0.2	27.1	54.9

[1] 0 = normal; 2 = previous history only; 3 = present symptoms only; 4 = positive physical examination without symptoms; 5 = positive X-ray without symptoms; 6 = positive X-ray or physical examination with present symptoms.

Table 6

PERCENT OF VARIANCE ASSOCIATED WITH EACH OF THE FIRST THREE PRINCIPAL COMPONENTS FOR 9 VARIABLES, BY SEX, AUGMENTED SAMPLE

NUMBER	VARIABLE	BOTH SEXES N = 252					MALE N = 109					FEMALE N = 143				
		Components					Components					Components				
		Total	I	II	III	Residual	Total	I	II	III	Residual	Total	I	II	III	Residual
	Total variance	100.0	46.4	17.0	11.1	25.5	100.0	42.4	15.5	11.2	30.9	100.0	43.7	13.5	13.4	29.4
1	Age	100.0	16.0	24.0	26.0	34.0	100.0	31.4	0.0	6.3	62.3	100.0	43.6	16.8	2.0	37.6
2	Height	100.0	65.6	13.7	2.3	18.4	100.0	44.9	24.0	0.6	30.5	100.0	44.9	17.6	13.7	23.8
3	Weight	100.0	32.5	14.4	11.6	41.5	100.0	30.3	16.0	21.2	32.5	100.0	2.9	74.0	0.0	23.1
6	Dynamometer	100.0	67.8	10.9	0.0	21.3	100.0	54.8	9.6	1.4	34.2	100.0	51.8	0.3	15.2	32.7
9	Puffmeter	100.0	62.4	7.8	1.0	28.8	100.0	50.4	21.1	10.2	18.3	100.0	63.3	0.0	4.4	32.3
14	$FEV_{1.0}$	100.0	81.0	4.0	0.5	14.5	100.0	74.0	6.3	3.6	16.1	100.0	74.1	0.0	5.3	20.6
16	VC	100.0	86.5	0.0	3.6	9.9	100.0	72.3	0.2	16.8	10.7	100.0	84.6	0.6	2.3	12.5
20	$FEV_{1.0\%}$	100.0	1.7	32.5	54.8	11.0	100.0	4.4	57.2	9.6	28.8	100.0	3.6	4.9	67.2	24.4
23	Category[1]	100.0	4.0	44.9	2.0	49.1	100.0	19.4	4.8	31.4	44.4	100.0	26.0	6.3	10.2	57.5

[1] 0 = normal; 2 = previous history only; 3 = present symptoms only; 4 = positive physical examination without symptoms; 5 = positive X-ray without symptoms; 6 = positive X-ray or physical examination with present symptoms.

Figure 1

PERCENT OF VARIANCE ASSOCIATED WITH EACH OF THE FIRST THREE PRINCIPAL COMPONENTS FOR TWENTY-THREE VARIABLES AND SUBSETS OF NINE VARIABLES, MALES AND FEMALES

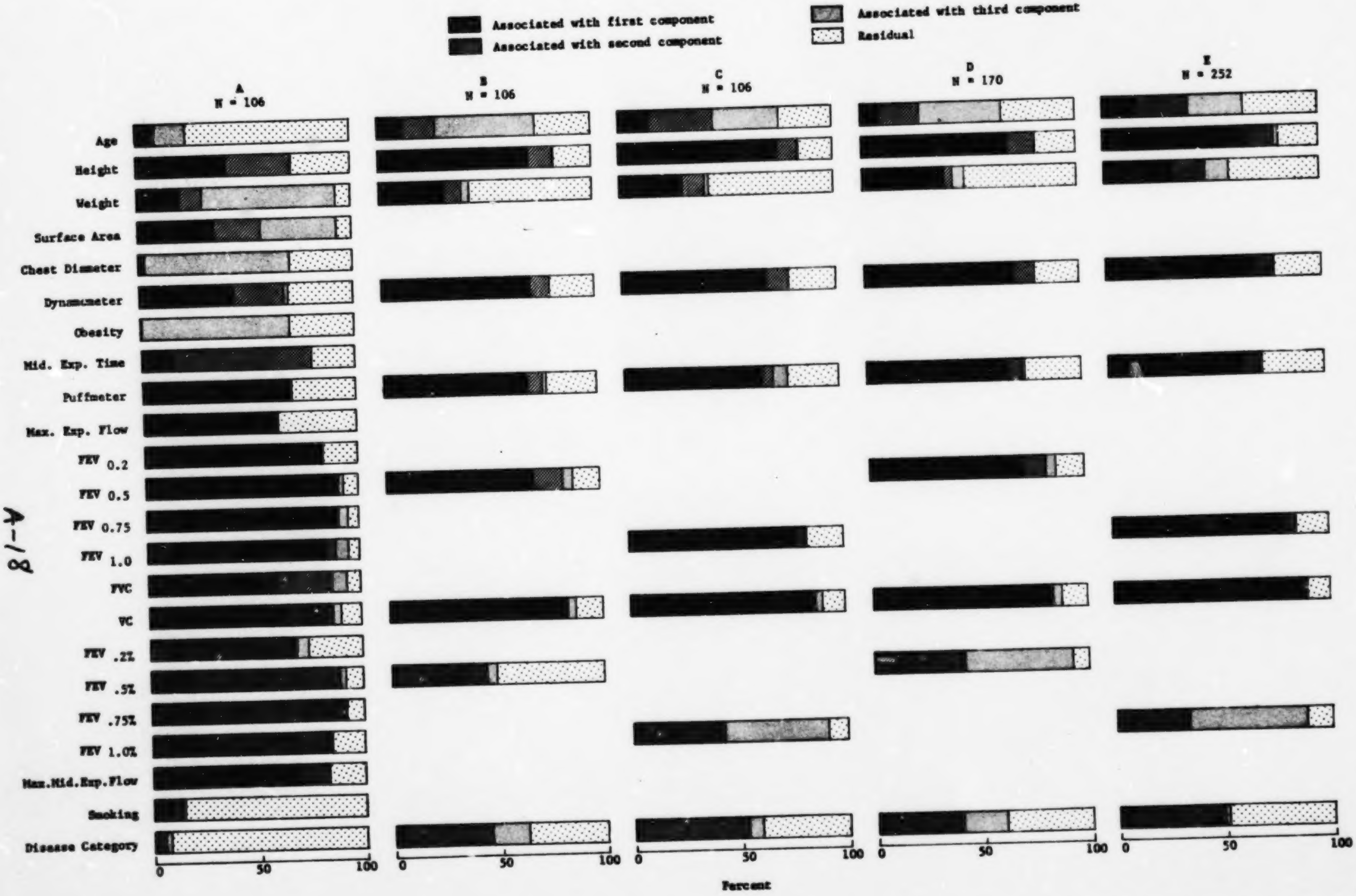


Figure 2

PERCENT OF VARIANCE ASSOCIATED WITH EACH OF THE FIRST THREE PRINCIPAL COMPONENTS FOR TWENTY-THREE VARIABLES AND SUBSETS OF NINE VARIABLES, MALES

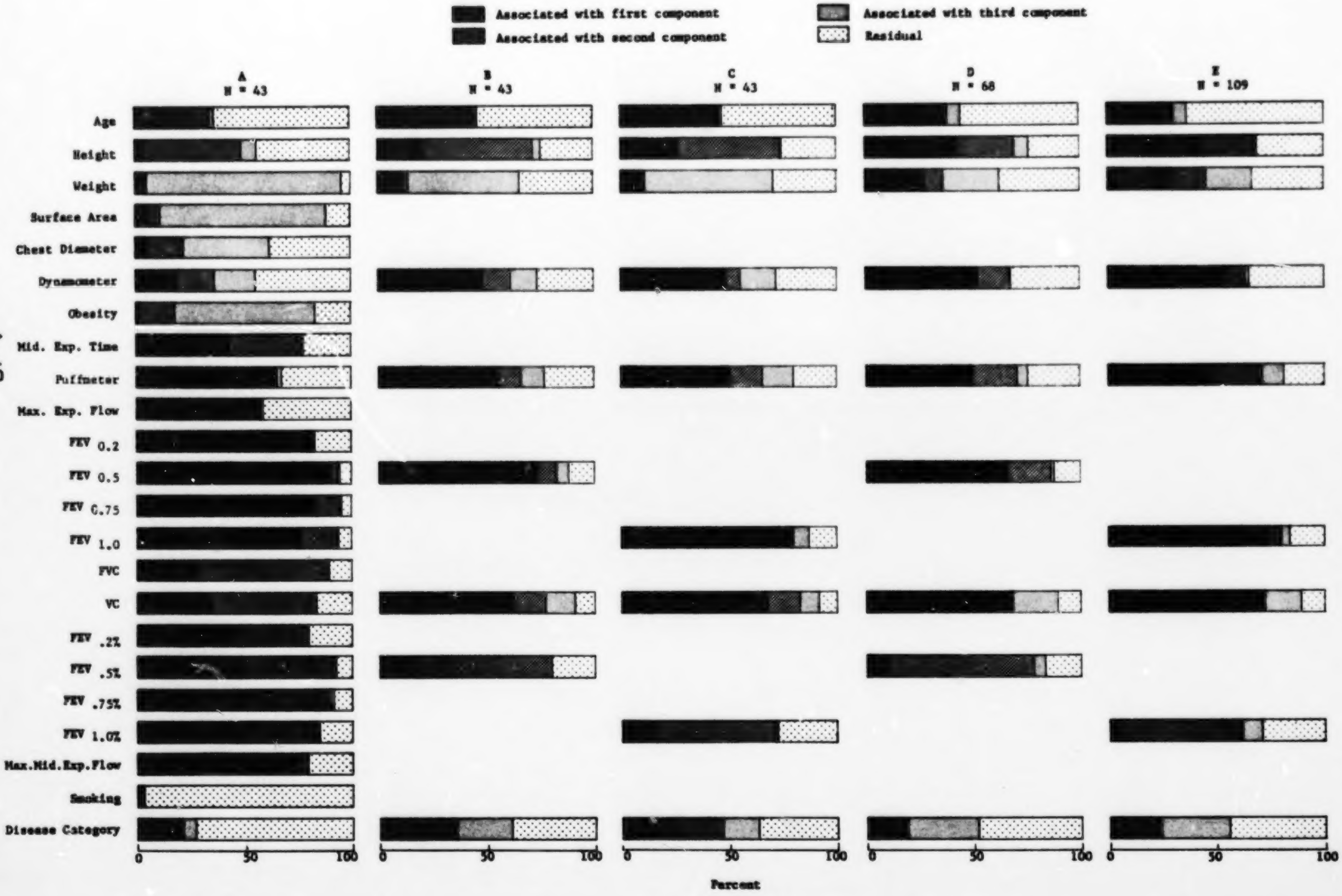
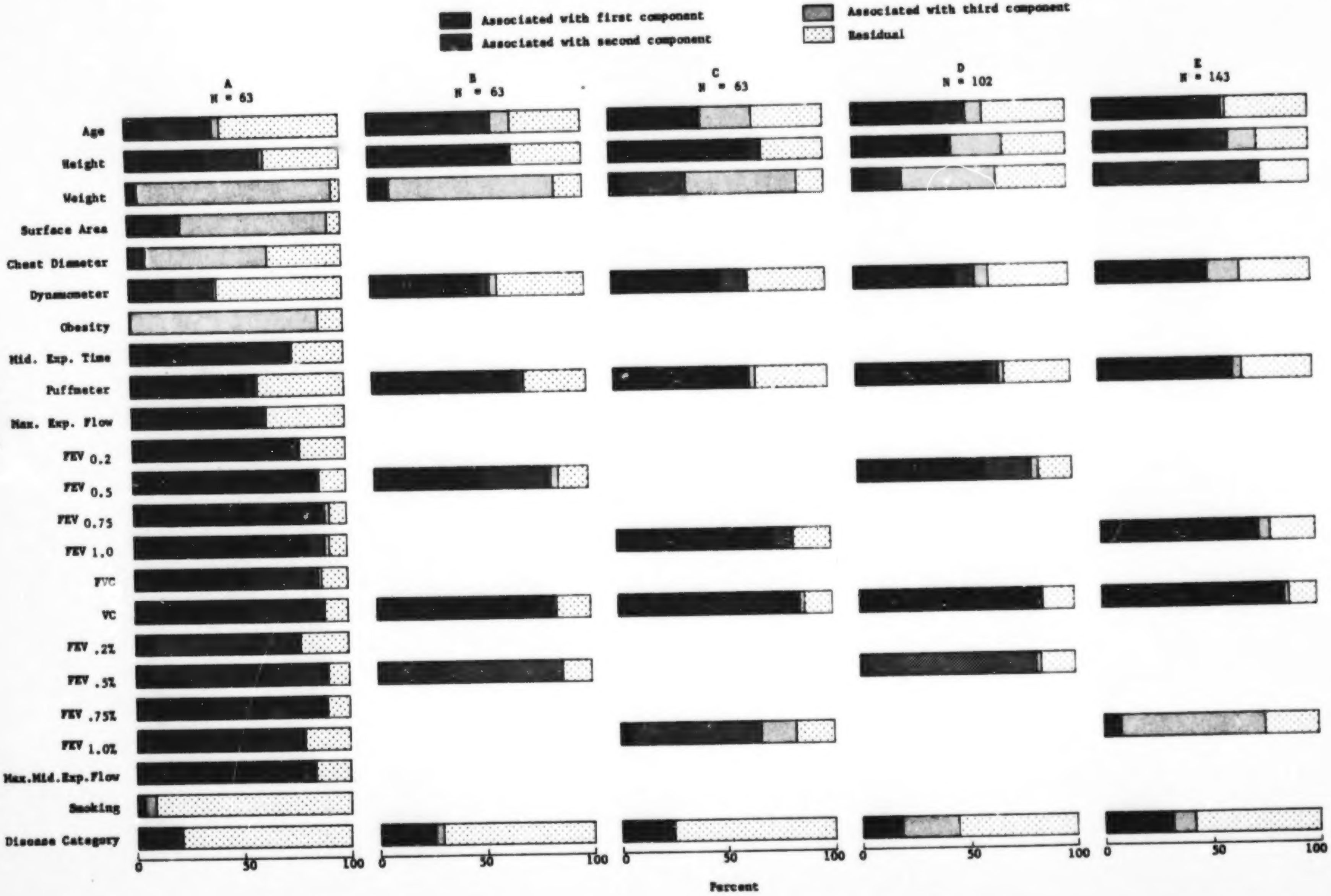


Figure 3

PERCENT OF VARIANCE ASSOCIATED WITH EACH OF THE FIRST THREE PRINCIPAL COMPONENTS FOR TWENTY-THREE VARIABLES AND SUBSETS OF NINE VARIABLES, FEMALES



A-21

Appendix A

CORRELATION MATRIX
ANTHROPOMETRIC, HEALTH, AND PULMONARY FUNCTION MEASUREMENTS

	NUMBER	VARIABLE	FEMALE N = 63																						
			NUMBER																						
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
			Age (years)	Height (inches)	Weight (pounds)	Body surface area (square meters)	Chest diameter (inches)	Dynamometer (strength of handgrip)	Obesity[1]	Mid-expiratory time (seconds)	Puffmeter (liters per minute)	Maximum expiratory flow (differentiated circuit - liters/minute)	0.2 second forced expiratory volume (liters)	0.5 second forced expiratory volume (liters)	0.75 second forced expiratory volume (liters)	1.0 second forced expiratory volume (liters)	Forced vital capacity (liters)	Spirometric vital capacity (liters)	0.2 second vol./forced vital capacity	0.5 second vol./forced vital capacity	0.75 second vol./forced vital capacity	1.0 second vol./forced vital capacity	Maximum mid-expiratory flow (liters per minute)	Smoking[2]	Respiratory disease category[3]
MALE N = 43	1	Age (years)		-37	00	-12	32	-50	11	-01	-41	-03	-08	-27	-39	-43	-45	-60	38	24	07	-04	-27	-24	27
	2	Height (inches)	-21		29	63	-09	61	-01	03	54	40	39	48	55	59	61	77	-12	-09	01	07	34	20	-25
	3	Weight (pounds)	-22	24		92	60	24	95	00	21	20	09	10	07	07	10	15	01	03	-03	-04	09	-12	-12
	4	Body surface area (square meters)	-24	62	91		46	43	77	-01	39	34	25	29	28	29	32	42	-01	03	01	01	23	-08	-20
	5	Chest diameter (inches)	26	-13	50	34		-07	64	-11	-05	07	-02	-09	-14	-17	-22	-19	21	20	14	10	-04	-21	14
	6	Dynamometer (strength of handgrip)	-40	41	40	50	-02		06	-03	52	23	24	31	38	39	41	58	-12	-09	01	03	33	05	-17
	7	Obesity[1]	-11	25	88	61	56	20		-02	05	08	-03	-04	-09	-10	-08	-08	04	06	-02	-05	-01	-20	-05
	8	Mid-expiratory time (seconds)	18	18	-06	04	03	-08	-14		-34	-38	-43	-42	-36	-29	04	-03	-57	-78	-89	-90	-77	13	07
	9	Puffmeter (liters per minute)	-33	23	05	13	-27	28	-06	-57		39	55	63	65	65	58	63	05	18	27	31	64	10	-29
	10	Maximum expiratory flow (differentiated circuit - liters/minute)	-31	11	08	10	-22	27	01	-39	50		84	73	64	58	46	42	58	56	52	43	59	06	-22
	11	0.2 second forced expiratory volume (liters)	-35	13	15	17	-30	40	07	-49	68	86		89	80	74	61	47	59	59	56	49	72	09	-27
	12	0.5 second forced expiratory volume (liters)	-47	23	11	18	-38	40	-01	-50	74	77	93		94	91	81	67	26	46	48	45	77	08	-33
	13	0.75 second forced expiratory volume (liters)	-52	31	11	21	-41	44	-05	-42	75	68	85	97		99	89	79	08	26	43	45	79	07	-39
	14	1.0 second forced expiratory volume (liters)	-53	38	10	23	-41	47	-09	-36	72	64	80	94	99		93	84	-04	14	32	37	75	07	-41
	15	Forced vital capacity (liters)	-43	50	01	22	-44	47	-24	13	45	47	56	70	80	85		86	-26	-14	-03	01	52	10	-42
	16	Spirometric vital capacity (liters)	-51	64	05	30	-35	53	-27	01	54	46	52	66	74	79	84		-28	-15	03	11	52	16	-35
	17	0.2 second vol./forced vital capacity	-09	-21	20	06	02	08	28	-68	48	69	76	58	41	32	-09	01		86	72	59	35	-04	06
	18	0.5 second vol./forced vital capacity	-16	-23	14	-01	01	00	23	-82	52	54	64	59	45	35	-14	00	91		88	78	54	-05	04
	19	0.75 second vol./forced vital capacity	-21	-20	16	03	01	01	24	-86	58	44	56	57	47	39	-14	04	82	97		87	71	-05	-04
	20	1.0 second vol./forced vital capacity	-22	-15	16	05	04	05	22	-87	55	38	50	53	45	38	-16	05	76	93	98		73	-06	-10
	21	Maximum mid-expiratory flow (liters per minute)	-31	09	15	15	-15	23	10	-79	76	55	73	79	76	70	33	37	61	68	72	71		00	-08
	22	Smoking[2]	-19	01	13	12	07	-03	13	-17	20	-07	07	11	14	13	08	03	01	05	08	08	27		18
	23	Respiratory disease category[3]	37	06	-25	-17	-07	-42	-27	28	-13	-36	-35	-39	-39	-38	-23	-15	-23	-27	-30	-30	-34	03	

[1] (weight of individual/mean weight for height and sex) x 100.

[2] 0 = nonsmoker; 1 = 20 or fewer cigarettes daily; 2 = 21-30 cigarettes daily; 3 = 31 or more cigarettes daily.

[3] 0 = normal; 2 = previous history only; 3 = present symptoms only; 4 = positive physical examination without symptoms; 5 = positive X-ray without symptoms; 6 = positive X-ray or physical examination with present symptoms.

--Discussion--

Using the figures in Appendix A, the correlation matrix, and Table I, which contains means and standard deviations for each variable, it is possible to express the relationship between any two variables in the form of a least squares linear regression equation. The formula to use is:

$$Y = a + b\ (X - \overline{X})$$

Where Y is the dependent variable or quantity to be estimated,

a is $\overline{Y}$, the sample mean from Table I,

b is $r_{xy} \cdot \frac{S_y}{S_x}$. Where S is the sample standard deviation from Table I and r_{xy} is the correlation between the two variables from Appendix A, and X is the independent variable with $\overline{X}$ shown in Table I.

As an example, to calculate the regression equation of puffmeter values on height for males, one would proceed as follows:

$$\text{Expected puffmeter value at any height} = \text{mean puffmeter} +$$
$$b(\text{height} - \text{mean height}) = 351.2 + .184 \times \frac{65.5}{2.5}\ (H - \overline{H}) =$$
$$351.2 + 4.82\ (H - 67.1) = 351.2 + 4.82H - 323.4 = 27.8 + 4.82\ H.$$

Another example would be to calculate for males the expected volume of air expired in 0.75 seconds given the volume expired in one second. The equation would be:

$$\begin{aligned} \text{Expected } FEV_{0.75} &= \text{Mean } FEV_{0.75} + b\ (FEV_{1.0} - \text{Mean } FEV_{1.0}) \\ &= 2.61 + .972\ \frac{0.60}{0.58}\ (FEV_{1.0} - 2.85) \\ &= 2.61 + 1.01\ (FEV_{1.0} - 2.85) \\ &= 2.61 + 1.01\ FEV_{1.0} - 2.88 \\ &= 1.01\ FEV_{1.0} - .27 \end{aligned}$$

Further application of the data in Appendix A and Table I permit calculation of partial and multiple correlation and regression equations. The reader is referred to J.P. Guilford: Fundamental Statistics in Psychology and Education, 3rd ed. McGraw Hill Book Co., Inc., New York, 1956, for formulae and examples.

1-23

APPENDIX III
BIMD 21
October, 1960

DIVISION OF BIOSTATISTICS
DEPARTMENT OF PREVENTIVE MEDICINE AND PUBLIC HEALTH
SCHOOL OF MEDICINE
UNIVERSITY OF CALIFORNIA AT LOS ANGELES

BIMD 21

PERIODIC REGRESSION

1. GENERAL DESCRIPTION:

a. This program performs Periodic or Harmonic Regression Analysis using the regression function of the form

$$Y = a_o + \sum_{i=1}^{n} \left[a_i \operatorname{Cos}(ic) + b_i \operatorname{Sin}(ic) \right] \quad \text{(nth harmonic)}$$

b. The program computes the periodic regressions successively up to the harmonic (n) specified in the Problem Card. It may stop well short of this, if a "good fit" is achieved with fewer terms.

c. The program computes and prints out an Analysis of Variance Table after each harmonic, successively.

d. For Analysis of Covariance, only a single covariate is allowed at one time; however, there is no limit to the number of covariates that can be processed singly. The number of harmonics in the Analysis of Covariance is the same as that determined by the direct analysis of the dependent variable.

e. The highest harmonic allowed in the program is 6.

f. The maximum number of sub-intervals of time is 400.

g. For the Analysis of Variance output an equal number of replications (between 2 and 20) must be present at each sub-interval of time.

h. Either a standard or a variable format for input data may be used.

i. The statistical procedure and a sample output of this program are shown in the Appendix.

2. CARD PREPARATION:

a. System Card

Necessary system card (I. D. etc.) will be stipulated at the IBM 704 (32K) or 709 installation at which the program is run. (Western Data Processing Center users should consult the WDPC Users Manual.) The format and meaning of cards that are input to this program are fixed and are described in this writeup.

b. Problem Card (One Problem Card for each problem)

Col.	1-6		Problem number or identification (It can be alphabetical characters.)
Col.	7, 8		Number of replications
Col.	9-11		Number of sub-intervals of time
Col.	12, 13		Scale Factor (See below)
Col.	14, 15		Transformation code (See below)
Col.	16-21		The value of a constant (c) in the form xxx. xx for 04 transformation ($\log_{10}(y + c)$). If 04 transformation is not used, leave these columns blank.
Col.	22, 23		Number of Covariance Problem Cards. (See the section e below.) Keypunch 00, if Analysis of Covariance is not desired.
Col.	24, 25	99	if a table of residuals is desired.
		00	if it is not desired.
Col.	26	n	The number of harmonic desired.
		0	Program specifies the number of harmonics.
Col.	71, 72	99	if a variable format is used.
		00	if it is not used.

A-25

Scale Factor (Only one scale factor is required for each problem.)

Since the program assumes all input data to be whole numbers (unless a decimal is punched), it may be necessary to scale input data in order not to lose significant digits or to place the decimal point. This is the purpose of the scale factor in the columns 12 and 13 of the Problem Card. All punched data are multiplied by 10 to a power equal to the scale factor.

Example:	Input Data	Data Keypunched	Scale Factor
1)	312	+00312	+0
2)	-31245000	-31245	+3
3)	.00031245	+31245	-8

Transformation Codes

The transformation code specified in the columns 14 and 15 applies to all input data. The transformation codes referred to are:

00	No transformation
01	$\sqrt{y}$ transformation
02	$\sqrt{y} + \sqrt{y+1}$ transformation
03	$\log_{10} y$ transformation
04	$\log_{10}(y + c)$ transformation When this transformation is desired, the value of c must be specified in Cols. 16-21 of the Problem Card.
05	$\sin^{-1}\sqrt{y}$ transformation
06	$\sin^{-1}\sqrt{y/(n+1)} + \sin^{-1}\sqrt{(y+1)/(n+1)}$ transformation
07	$1/y$ transformation

c. Variable Format Card

If input data cards have been prepared using a format different from that specified in the next section (or if another format is desired for other reasons) a variable format may be specified. If a variable format is to be used, 99 must be punched in appropriate columns of the Problem Card and the Variable Format Card must immediately precede the first Input Data Card. The format specification starts in the first card column and must be completed in the first 72 columns of this one card. The word FORMAT is not used. For example,

(12F3.0, F4.0, 11F2.0)

punched in the first 20 columns of the Variable Format Card will describe 12 three column fields, 1 four column field, and 11 two column fields. Another example,

(5X, 2F6.0, F1.0, 3X, F5.0/5X, F6.0)

punched in the first 31 columns of the Variable Format Card will direct entry of five data fields as follows:

1) 6 digit field in cols. 6 to 11 of first card
2) 6 digit field in cols. 12 to 17 of first card
3) 1 digit field in col. 18 of first card
4) 5 digit field in cols. 22 to 26 of first card
5) 6 digit field in cols. 6 to 11 of second card

Thus the data are on two cards for each case.

If the Problem Card does not have 99 punch calling for a variable format, the program will receive data prepared according to the standard format explained in the following section.

d. Input Data Cards. (Keypunch data row-wise)

The form of input data is illustrated in the following example.

	Sub-intervals				
Replications	t_o	t_1	t_2	t_3	etc.
1)	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	.
2)	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	.
3)	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	.
etc.	.	.	.	.	.

The preparation of the Input Data Cards is as follows:

Col.	1-6	A sign and $y_{1,1}$ value
Col.	7-12	A sign and $y_{1,2}$ value
Col.	13-18	A sign and $y_{1,3}$ value

etc.

If there are more than 12 sub-intervals, continue keypunching a second card in the same manner. The maximum number of sub-intervals is 400.

Leading zeros may be indicated by blanks or by punched zeros. Positive signs need not be punched.

Each replication must begin on a new card:

$y_{2,1}, y_{2,2}, y_{2,3}, \ldots,$ etc.

e. Covariance Problem Card (One card for each covariate)

This card and the covariate data cards are prepared only when a non-zero value is punched in Col. 22 and 23 of the Problem Card.

The preparation of the Covariance Problem Card is as follows:

Col.	1-6	Covariate number or identification (It can be alphabetical characters.)
Col.	7, 8	Scale Factor
Col.	9, 10	Transformation Code
Col.	11-16	The value of a constant (c) in the form xxx.xx for $\log_{10}(y + c)$ transformation. If this transformation is not used, leave these columns blank.
Col.	71, 72	99 if a variable format is used 00 if it is not used.

The Scale Factor, Transformation Code, Variable Format Card and Input Data Cards are indicated and/or prepared in the same manner as those described for the variate.

There is no limit to the number of different sets of covariate data. Each Analysis of Covariance is performed separately with the original variate. The program expects as many Analysis of Covariance as the number specified in Col. 22 and 23 of the Problem Card.

g. Finish Card

This card will notify the program that the entire job is finished. This card has the following format:

Col. 1-6 FINISH

4. ORDER OF CARDS:

More than one problem can be processed consecutively. The following example illustrates the order of cards:

.. System Cards
.. and/or
.. Program Cards
1) Problem Card for the first problem
2) Variable Format Card (if used)
3) Input Data Cards for variate
4) First Covariance Problem Card (if used)
5) Variable Format Card (if used)
6) Input Data Cards for covariate
7) Second Covariance Problem Card (if used)
..
..
.. Last Covariance Problem Card (if used)
.. Variable Format Card (if used)
.. Input Data Cards for covariate
..
..
.. Problem Card for the last problem
.. Variate Format Card (if used)
.. Input Data Cards for variate
.. First Covariance Problem Card (if used)
..
..
.. Finish Card

5. APPENDIX:

The statistical model requires that we assume y_{rt} was observed at $x_t = 2\pi t/k$. Harmonic Analysis of this series consists in the fitting by least-squares the regression function

$$Y_t = a_o + \sum_{j=1}^{n} [a_j \text{ Cos } (j\ t\ c) + b_j \text{ Sin } (j\ t\ c)]$$

having 2 n + 1 regression coefficients

$$a_0, a_1, a_2, \ldots, a_n;\ b_1, b_2, \ldots, b_n$$

where $c = 2\pi/k$, converting the t to radian measure.

A constant number "m" of observations y_{rt} $(r = 1, 2, 3, \ldots, m)$ are taken at each value of x_t $(t = 0, 1, 2, \ldots, k-1)$. These observations may actually occur at the same point "t" in time or may be observations at corresponding points on different cycles of the process.

In the following steps the indicated computations are carried out. The values computed from Steps 2 (b), 3, 4, 5, 6, and 7 are printed as part of the output of this program.

STEP 1.

(a) Sine and Cosine Functions

$$u_{tj} = \text{Cos } (c\ t\ j)$$

$$v_{tj} = \text{Sin } (c\ t\ j)$$

$$D_{uj} = \sum_{t=0}^{k-1} u_{tj}^2$$

$$D_{vj} = \sum_{t=0}^{k-1} v_{tj}^2$$

(b) Sums

$$T_r = \sum_{t=1}^{k} y_{rt}$$

$$T_t = \sum_{r=1}^{m} y_{rt}$$

$$\text{TOTAL} = \sum_{r=1}^{m} \sum_{t=1}^{k} y_{rt}$$

$$\text{SUM 2} = \sum_{r=1}^{m} \sum_{t=1}^{k} y_{rt}^2$$

$j = 1, 2, \ldots, n$; where n is the number of harmonics

STEP 2 Regression Coefficients

(a) For each replicate	(b) Average for all replicates
$a'_{rj} = \sum_{t=0}^{k-1} u_{tj} y_{rt}$	$A'_j = \sum_{t=0}^{k-1} u_{tj} T_t$
$b'_{rj} = \sum_{t=0}^{k-1} v_{tj} y_{rt}$	$B'_j = \sum_{t=0}^{k-1} v_{tj} T_t$
$a_{rj} = a'_{rj} / D_{uj}$	$A_j = A'_j / mD_{uj}$
$b_{rj} = b'_{rj} / D_{vj}$	$B_j = B'_j / mD_{vj}$
$a_{ro} = T_r / k$	$A_o = \text{TOTAL} / mk$

STEP 3

The predicted response at time t using the average fitted curve,

$$Y_t = A_o + \sum_{j=1}^{n} [A_j u_{tj} + B_j v_{tj}]$$

the actual (mean) response at time t,

$$\bar{y}_t = T_t / m$$

and the residuals (difference),

$$d_t = Y_t - \bar{y}_t$$

STEP 4 Analysis of Variance

Sums of Squares	Degrees of Freedom
(a) Correction for mean	
$CM = (\text{TOTAL})^2/m\,k = S_7$	1
(b) Between replicates	
$S_1 = \sum_{r=1}^{m} T_r^2/k - CM$	m - 1
(c) Harmonic Intensities $(a_i + b_i)$**	
$S_{2j} = A'^2_j + B'^2_j$	2
(d) Scatter about curve	
$S_3 = \sum_{t=1}^{k} T_t^2/m - CM - \sum_{j=1}^{n} S_{2j}$	k - 2n - 1
(e) Interaction of Replicates X Intensities	
$S_{4j} = \sum_{r=1}^{m} [a'^2_{rj} + b'^2_{rj}] - S_{2j}$	2 (m - 1)
(f) Total Sum of Squares	
$S_6 = \text{SUM2} - CM$	mk - 1
(g) Interaction of Replicate X Scatter, by difference,	
$S_5 = \text{SUM2} + CM - [1/k \sum_{r=1}^{m} T_r^2 + 1/m \sum_{t=1}^{k} T_t^2 + \sum_{j=1}^{n} S_{4j}]$	(m - 1)(k - 2n - 1)

**Harmonic Intensities $(a_i + b_i)$ is used here to designate the sum of the effects of the harmonic coefficients, not the sum of the coefficients directly.

(h) Interaction of Replicate X Amplitude

$$S_{8j} = 2\,S_{2j} + S_{4j} - 2\sqrt{S_{2j}/m}\;\; R_j \qquad\qquad m - 1$$

$$\text{where } R_j = \sum_{r=1}^{m} \sqrt{a'^2_{rj} + b'^2_{rj}}$$

(i) Interaction of Replicate X Phase

$$S_{9j} = S_{4j} - S_{8j} \qquad\qquad m - 1$$

STEP 6

The above Sums of Squares, appropriate for F ratios, are divided by their respected degrees of freedom

$$MS_{p(j)} = S_{p(j)} / DF, \qquad\qquad p = 1, 2, 3, 4, 8, 9$$

STEP 7 Test of significance (F Ratios)

For a discussion of the proper tests of significance see Bliss, C. I. *.

STEP 8

The Range, Semi-Amplitude, Phase Angle, time of maximum response, time of minimum response are computed.

STEP 9 Analysis of Covariance

For the sums of squares of the covariate, X_{rt}, all computations being identical except for sums of products (xy) in which the numbers formerly squared are here cross-multiplied.

---

*Reference: Bliss, C. I., (1958), Periodic Regression in Biology and Climatology; Bulletin 615; the Conn. Agricultural Experimental Station; New Haven, 11, Conn.

```
BIMD 021
UCLA MED CENTER                                                    BIMD 21

JOB CODE HUMID

NUMBER OF REPLICATIONS  3
NUMBER OF SUBINTERVALS    24
TRANSFORMATION CODE  0

            TABLE OF RESIDUALS
          FOR HARMONIC NUMBER 1

 TIME    ACTUAL     PREDICTED   RESIDUAL

    0     0.4473       0.4741    -0.0268
    1     0.4487       0.4858    -0.0371
    2     0.5080       0.4969     0.0111
    3     0.5003       0.5067    -0.0064
    4     0.5097       0.5146    -0.0049
    5     0.5357       0.5200     0.0157
    6     0.5433       0.5225     0.0209
    7     0.5443       0.5220     0.0224
    8     0.5140       0.5184    -0.0044
    9     0.4997       0.5122    -0.0125
   10     0.5013       0.5036    -0.0022
   11     0.4910       0.4932    -0.0022
   12     0.4870       0.4818     0.0052
   13     0.4083       0.4702    -0.0618
   14     0.4763       0.4590     0.0173
   15     0.4440       0.4492    -0.0052
   16     0.4840       0.4413     0.0427
   17     0.4610       0.4360     0.0250
   18     0.4123       0.4334    -0.0211
   19     0.4303       0.4340    -0.0036
   20     0.4200       0.4375    -0.0175
   21     0.4390       0.4438    -0.0048
   22     0.4957       0.4524     0.0433
   23     0.4697       0.4627     0.0070

PARAMETER ESTIMATES

CONSTANT(MEAN)=          0.4780

COEFFICIENTS
A(1)=      -0.0039  B(1)=       0.0445

SEMI-AMPLITUDE =                 0.0447
PHASE ANGLE (RADIANS)=           1.6571
RANGE =                          0.0894
TIME OF MAXIMUM RESPONSE =       6.3296
TIME OF MINIMUM RESPONSE =      18.3296
```

```
                                   ANALYSIS OF VARIANCE
SOURCE CF VARIATION   DF   SUM OF SQUARES  MEAN SQUARE      F RATIOS    DF FOR F TESTS
*************************************************************************************
BETWEEN REPLICATES      2          0.0099        0.0049        6.5410    (   2,    42)
EFFECT CF (A1+B1)       2          0.0719        0.0359        4.9200    (   2,    5.44)
SCATTER ABOUT CURVE    21          0.0389        0.0019        2.4561    (  21,    42)
REPLICATE X (A1+B1)     4          0.0248        0.0062        8.2180    (   4,    42)
REPLICATE X SCATTER    42          0.0317        0.0008
*************************************************************************************
TOTAL                  71          0.1773
CORRECTION TO MEAN      1         16.4480
*************************************************************************************
REPLICATE X AMPLI 1     2          0.0229        0.0114       15.1359    (   2,    42)
REPLICATE X PHASE 1     2          0.0020        0.0010        1.3002    (   2,    42)
*************************************************************************************
```

```
COVARIANCE CODE  HEAT

TRANSFORMATION CODE  0

                                          ANALYSIS OF COVARIANCE

SOURCE OF VARIATION     DF       XX              XY            YY     *      BS2-BE2   DF          RED.MS      F RATIOS
*********************************************************************************************************************
                                                                              0.0079    2          0.0010       2.336
BETWEEN REPLICATES       2     1.4603          0.1123         0.0099  *       0.0050    2          0.0374       5.850*
EFFECT OF (A1+B1)        2     0.0476          0.0241         0.0719  *       0.0031   21          0.0017       4.0213
SCATTER ABOUT CURVE     21     1.4652          0.0847         0.0389  *       0.0075    4          0.0043      10.1890
REPLICATE X (A1+B1)      4     0.1988          0.0434         0.0248  *       0.0143   41          0.0004      33.9642
ERROR                   42     1.1675          0.1292         0.0317  *
****************************************************************************
ERROR+REPLICATES        44     2.6278          0.2415
ERROR+(A1+B1)           44     1.2151          0.1533                      *EST DF FOR DENOMINATOR 1 ABOVE=  6.52
ERROR+SCATTER           63     2.6328          0.2140                       REGRESSION COEFFICIENT B(YX)=  0.1107
ERROR+REPS X (A1+B1)    46     1.3664          0.1726
*********************************************************************************************************************

                                                                           BE2=(XY).(XY)/XX     FROM ERROR ROW
                                                                           BS2=(XY).(XY)/XX     FROM SUMS IN LAST ROWS
                                                                           RED.MS=(YY-(BS2-BE2))/DF
```

ACKNOWLEDGMENT

The individuals herein listed constituted the staff of this project. Their ideas and other contributions to its operation are gratefully acknowledged.

California State Department of Public Health
Berkeley, California

John R. Goldsmith, M.D.
Carl J. Hopkins, Ph.D., Special Consultant
H.H. Hechter, Associate Public Health Analyst
Margaret Deane, Associate Public Health Analyst

Norman Perkins, Assistant Research Technician
Dorothy Tuma, Junior Public Health Analyst
Michael Zwerdling, Junior Public Health Analyst
Lynn Hayward, Graduate Student

LeMar Hoaglin, Administrative Assistant
Rose Jackson, Typist
Helen Kaiser, Clerk
Signorina Tucker, Typist

Professor Frank J. Massey, Jr., Ph.D.
Associate Professor of Biostatistics
University of California
Los Angeles, California

A-36